AFRL-IF-RS-TR-2005-141
Final Technical Report
April 2005

CYBERSPACE ASSURANCE METRICS: UTILIZING MODELS OF NETWORKS, COMPLEX SYSTEMS THEORY, MULTIDIMENSIONAL WAVELET ANALYSIS, AND GENERALIZED ENTROPY MEASURES

University of South Carolina Research Foundation

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

STINFO FINAL REPORT

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

AFRL-IF-RS-TR-2005-141 has been reviewed and is approved for publication

APPROVED: /s/
JAMES L. SIDORAN
Project Engineer

FOR THE DIRECTOR: /s/
WARREN H. DEBANY, JR.
Technical Advisor
Information Grid Division
Information Directorate

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)	**2. REPORT DATE** April 2005	**3. REPORT TYPE AND DATES COVERED** **Final Sep 02 – Mar 04**

4. TITLE AND SUBTITLE

CYBERSPACE ASSURANCE METRICS: UTILIZING MODELS OF NETWORKS, COMPLEX SYSTEMS THEORY, MULTIDIMENSIONAL WAVELET ANALYSIS, AND GENERALIZED ENTROPY MEASURES

5. FUNDING NUMBERS

G - F30602-02-1-0231
PE - 61101E
PR - PO91
TA - A1
WU - 06

6. AUTHOR(S)

Joseph E. Johnson
Vladimir Gudkov

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of South Carolina Research Foundation
901 Sumter Street, Suite 511
Columbia SC 29208

8. PERFORMING ORGANIZATION REPORT NUMBER

N/A

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFRL/IFGB
525 Brooks Road
Rome NY 13441-4505

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

AFRL-IF-RS-TR-2005-141

11. SUPPLEMENTARY NOTES

AFRL Project Engineer: James L. Sidoran/IFGB/(315) 330-3174 James.Sidoran@rl.af.mil

12a. DISTRIBUTION / AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

The problem is addressed of developing a very general mathematical foundation for networks that permits practical application in the monitoring of large networks such as the internet for both known and unknown attacks, intrusions, worms, viruses, and generally for destructive agents and processes. The PI, under the funding of this grant, has discovered a strong connection between the topological specification of a network in the form of a connection matrix and the branches of mathematics known as continuous group theory and Markov processes. Based upon this research he has proposed that entropy metrics, and the associated cluster analysis of the network so measured by these metrics, can be useful indicators of aberrant processes and behavior. Other team members have obtained important connections using higher order Renyi entropy metrics, and complexity theory to both monitor real networks and to study networks by simulation.

14. SUBJECT TERMS

Network security, entropy metrics, cluster analysis, complexity theory

15. NUMBER OF PAGES 89

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Table of Contents

List of Appendices

1 Introduction and Overview

Currently there is no general foundational theory for networks that can reduce the extremely large (trillions of real values for a million node network) into a few characteristic values (called network metrics) that distill the essential aspects of the network into just a few metrics (functions of the network matrix values). It is well known that a network is characterized exactly by the n^2-n non-negative off-diagonal elements of the connection matrix whose elements C_{ij} consist of the 'strength' of the connection between nodes i and j. Our objective is to develop a very general mathematical foundation for networks that permits practical application in the monitoring of large networks, such as the internet, for both known and unknown attacks, intrusions, worms, viruses, and generally for destructive agents, processes, and system failures.

This document describes the final technical results for *Cyberspace Assurance Metrics: Utilizing Models of Networks, Complex Systems Theory, Multidimensional Wavelet Analysis and Generalized Entropy Measures*, a project sponsored by the DARPA Information Assurance and Survivability Program, and funded through the United States Air Force Research Laboratory. The researchers, led by Dr. Joseph E. Johnson, have discovered a strong connection between the topological specification of a network in the form of a connection matrix and the branches of mathematics known as continuous group theory and Markov processes. Based upon this research we have proposed that entropy metrics, and the sub-networks arising from an associated cluster analysis of the network so measured by these metrics, can be useful indicators of aberrant processes and behavior. This is achieved by tracking the generalized entropy metrics and identifying the normal range of values. Our hypothesis is that as abnormal ranges are observed on the entropy values of a network or sub-networks, there is evidence that the entropy metrics for that subnet go to abnormal values as the topology changes in an abnormal way. Other team members have obtained important connections using higher order Renyi entropy metrics, and complexity theory to both monitor real networks and to study networks by simulation.

This report is organized into four sections: Section One discusses the problem and potential approaches for addressing network metrics; Sections Two and Three discuss the technical and mathematical approaches taken by the researchers, especially the mathematical equations and reasoning used; and Section Four addresses the results and conclusions of the project with a discussion of future research potential. The Appendices are technical papers written and published by the researchers during the grant cycle.

1.1 Statement of the Problem

Systems, such as a gas, can be described exactly, but uselessly, if one could specify the three coordinates and the three momenta of each of the 10^{24} particles that make up the gas. The theory of thermodynamics reduces this astronomical number of variables to just a few holistic variables which are extremely useful for the description of the system as a whole: temperature, pressure, entropy, volume, etc. These variables are intuitive, and hierarchical (in the sense that they can be applied to a spatial sub domain of the system). The internet and other types of large networks are other examples of systems with a vast number of coordinates that are specified by the type of connection (matrix) between one node (or point) in the system and another node. For example C_{ij} might represent just whether a connection exists from i to j (C_{ij} =1) or not (C_{ij} =0), or it could represent the extent of the connection (as a non-negative real number). For a network of a million nodes (10^6) one then has a trillion numerical values (10^{12}) all of which are constantly changing as new connections are made and others are broken with $C_{ij}(t)$ as a matrix function of time. As opposed to the science of thermodynamics and statistical mechanics, there is no foundational intuitive, hierarchical, set of metrics (functions of C_{ij}) for the description of networks such as we have with thermodynamics or statistical mechanics. Heat and temperature require the concept of approach to equilibrium and there is no obvious way to characterize internet traffic as approaching equilibrium that supports reasonable concepts of equilibrium. As there is no meaning associated with distance in a network, the concepts of pressure and volume also do not have a parallel definition from thermodynamics.

1.2 Technical Approaches

We have found that entropy (as a measure of system disorder) is a good candidate for a network metric as shown below. In what follows we can speak equivalently of Information which is defined as the negative of entropy, and representing system order. This project on network security addressed the problem of the identification of attacks and intrusions in networks such as the internet. As there is no general mathematical theory providing an underlying structure for networks, past efforts have centered on detection of specific types of attacks, intrusions, worms, viruses, and malevolent processes. We have sought general techniques and specifically a mathematical foundation that will provide a set of metrics that describe holistic aspects of the vast detail of large network topology. The author has been able to link the connection matrix that exactly describes a network, to a member of a Lie algebra of generators of continuous Markov processes. Thus in one step we can use mathematical results and insights in multiple branches of mathematics to each shed light on the other: network theory, Markov theory, continuous group theory, diffusion theory (increasing entropy), and information theory. This work suggested also that the metric of entropy as a measure of disorder on the network would be a meaningful measure of the network and its subnets. Specifically, we were able to show that the entropy measures the rate of diffusion to a state of disorder of a fluid that flows on that same network topology thus providing an intuitive foundation for the entropy metrics. Finally, it can be shown that the entropy on a network is related to the degree of clustering in the topology. The entropy metrics are therefore suggested as good monitors for the changes in topology over time with the assumption that the connectivity and flow rates of subnets which are under attack will change substantially and since the entropy metrics is defined by the order or disorder of this topological structure, then these metrics should reflect deviations from their normal patterns of temporal change thus indicating attacks and intrusions.

Dr. Vladimir Gudkov, working with the author, made very substantial progress in the understanding of various higher order Renyi entropies and their intuitive interpretation as measuring order and disorder in the higher dimensional aspects of network connectivity. He also was able to run a substantial number of simulations and there showed that the differences of different generalized entropies are highly sensitive to cluster formation and thus to topological change in the network. His studies in complexity theory related to networks indicate that the use of higher order Renyi entropy metrics can be rapidly computed on real networks. His work on simulated networks was extremely enlightening in the ability to monitor the various types of generalized entropy under diverse formations and dissolutions of topological clustering. In this work he led graduate students in the development of new lighter weight SNORT type programs for the capturing of internet traffic which could in turn be used to compose an associated network matrix. Subsequently their work on relatively small networks at USC was able to compute and track entropy variations over time. These results are reported in the attached papers by Gudkov and Johnson which have been submitted to the Los Alamos Preprint Library.

Working with Gudkov and Johnson, Dr. Shmuel Nussinov was able to find a very revolutionary way to identify clusters in a large network. One can show that just as three nodes can be equally spaced at the points of an equilateral triangle in two dimensions, and as four nodes can be equally spaced in three dimensions at the points of a tetrahedron, then the n nodes of a network can be equally space by placing them at equidistant points in an n-1 dimensional space. A system using forces that pull nodes closer if they are connected and pushing them apart if they are not connected provides equations that dynamically iterate the condensation of clusters out of the network.

2 Methods and Procedures

The following section discusses the technical assumptions and procedures used by the researchers.

2.1 Connection Matrix

The connection matrix $C_{ij}(t)$, as defined for a set of network flows among nodes, contains the entirety of information of that structure and further represent as a sequence of representing their change over time. Such a connection matrix changes over time, but observed directly, these million of microscopic changes do not provide the macro-scale variables, or network metrics, that could be used to monitor the overall state of a network or its sub-networks. Generalized entropy functions provide a set of metrics that directly depend upon the clustering densities of nodes and which ‘summarize the topology’ and thus afford potential intuitive macro-scale metrics to observe the $C_{ij}(t)$ dynamical changes. One of the hypotheses of our general approach is that abnormal changes in these entropy metrics will reflect abnormal structural changes in the connection matrix indicative of intrusions and abnormal behavior, and yet not be sensitive to uninteresting small statistical fluctuations. Thus insights into the interpretation and meaning of these generalized entropy type functions and the identifications of subnets are important to guide the practical computational applications.

Any connection matrix is defined only by (nonnegative) off-diagonal elements and thus the diagonals admit any arbitrary value. We have been able to show that the choice of diagonal elements to be the negative of the sum of the row elements, for each column, is the generator of a Markov transformation of flows on the network via M(s) = $\exp^{sC_{ij}}$ where s is a parameter generating the extent of the flow. The exponentiation of a (square) matrix is defined in the same manner as $e^x = 1 + x + x^2/2! \ldots$ As this is a continuous Markov transformation, it represents the flow of a conserved quantity which is the sum of the values of the vector upon which the transformation acts. Formally one can show that all allowable C matrices with this diagonal are members of a Lie algebra that generate Markov type Lie group flows. The technical aspects of this work are attached as a separate paper. Such flows are not to be interpreted necessarily as the flow of information but rather the flows reflect the topological structure, connectivity, and clustering. Other choices of diagonal values can be shown to give continuous general linear transformations that provide for source and sink (of a hypothetical fluid such as information or money) at each node. This connection between network connection matrices and the rich mathematical domain of continuous (Lie) groups and algebras can provide an intuitive guide for the entropy dynamics of a network. In fact the generalized entropy equations of the last section here can be seen as the measure of the generalized entropy increase when a conserved fluid begins at a given node and is dispersed by the topology in first order. Thus the important advantage of the link with continuous Markov transformations is that one obtains an extensive mathematical insight from Lie group theory that can provide guidance in the choice of entropy functions and where they should be calculated.

2.2 Dynamical Network Changes

One must be careful not to confuse the dynamical flow of the Markov conserved 'fluid' (which is used as a guide to network entropy calculations) with the dynamical changes in the network which arise from the time dependence of $C_{ij}(t)$. In the practical world, the $C_{ij}(t)$ is defined over a time collection period centered about the instantaneous time t. As clusters form and dissolve, and topological structures change with time due to the dissolution and reforming of different connections, we need to track the changes in the topology over time for some sub-net or the entire network. Any natural method of identification of a sub-network, c, can be used for the definition of a domain for the calculation of the generalized entropy over time $R_q(c,t)$ which serves as a macro level variable that should be insensitive to small changes in the topology of the sub-network c. Thus one can monitor the local entropy densities over networks and sub-networks, c, which can be tracked over time as a density at a specific node or for any cluster as the cluster dissolves over time under the changes in $C_{ij}(t)$. The use of other spectral 'classifications' of nodes can also be used to group nodes for entropy monitoring including eigenvectors and groups of nodes defined from natural network design configurations. One such example of definition is obtained by classifying nodes into classes depending first on the number of connections to other nodes. Then to separate the subclasses of each class depending on the number of connections to other classes etc. until the procedure is exhausted.

We have also been able to show that for the second order Renyi entropy, an associated information function can be defined that consists of a natural expansion in a Taylor series about the evolution parameter, s, which expansion has terms which are the diagonals of the various powers of the C_{ij} matrix, terms which are familiar to those in the field as useful expressions of the structure. We have additionally been able to show that the powers of C_{ij} (with diagonals set to zero) where the diagonal at each stage is set equal to zero prior to multiplying by the next power of C, give a set of diagonal values which count the non-recurring paths through the starting nodes. Such powers are not linearly related to the original C matrix and have other interesting properties.

3 Mathematical Background

This section describes the mathematical equations and theory that are the foundation of this work.

3.1 Network Topology

The traditional representation of an undirected graph or network topology utilizes an 'adjacency matrix', C, where $C_{ij} = 1$ if nodes i and j are connected and 0 otherwise, thus leaving the diagonal assignments arbitrary. We first show that with the assignment of the diagonal values to be the negative of the sum of the non-diagonal elements in each row, C becomes an element of a (Markov type) Lie algebra that generates conserved flows on that network. This provides a unique connection between static network topologies and flows generated by the associated Lie group (or monoid if non-negative components are used). Thus a topology specified by C generates irreversible continuous Markov transformations (flows) along the connections in that topology that represent diffusion and thus increasing entropy of the conserved entity and achieving a single final equilibrium state. Next, using a second order Renyi (generalized Shannon) entropy metric, we show that the diagonal values of the powers, C^n , (often utilized to study the topology of a network) are the n^{th} derivatives evaluated at t = 0 of a function of the Renyi entropy of the flowing entity. Thus the diagonals of the powers of C can often provide a spectral ordering of nodes often with symmetry broken with higher powers, and offering a 'series expansion' of a network. We next show that the counting of non-returning paths of k steps between two nodes is given by powers of C with the diagonal removed at each power. Fourthly, the view of a topology is explored in relation to flows and diffusion associated with these transformation groups and extensions to the general linear group for arbitrary diagonal C values using the n different diagonal Abelian transformations. It is shown that (diagonal) values contained in the derivatives of the information function at t=0 contain all off diagonal information thus suggesting that the eigenvalues of these n different C matrices could be used to more extensively classify network topologies. These transformations can be interpreted as supplying extra 'fluid' at any arbitrary node during the process of approach to equilibrium. The diagonal terms and eigenvalues here explored provide metrics for monitoring topologies and topological changes in networks along with an intuitive model of flows of a conserved fluid by diffusion toward equilibrium. These metrics are hoped to provide a deeper understanding of network topologies and structures which could be useful for problems both in solid state physics, internet dynamics and intrusions, and other network applications in social, engineering, and economic problems.

3.2 An Understanding of Networks

The extensive mathematical literature on graphs (networks) has seen substantial multidisciplinary research activity and interest over the last half century centering on problems in social, transportation, organizational, utility, electrical circuit, and financial networks in addition to those in physics. But most recently it is recognized that a better understanding of networks is critical because of our reliance on the internet, an incredibly large and complex communication network that is rapidly becoming totally critical for modern society.

Since a network (or graph) is defined as a set of points called nodes with lines connecting some but generally not all of the nodes. Sets of nodes that are disconnected from the rest of the structure can be removed and treated as a separate network leaving networks for which each node is connected to at least one other node in the topology. If the lines connecting the nodes are unidirectional then the network is said to be undirected and if the connections are directed then the network is said to be directed. The problem is similar to the numbering of identical particles in quantum theory but here one cannot perform a 'second quantization' because the underlying topological connections provide some nodal distinctions.

A number of researchers have independently suggested that the eigenvalues of the connectivity matrix (by any of the three methods of assigning the diagonal discussed above) will have values, which are in one to one (isomorphic) correspondence for, and only for, topologically identical networks. This would have been a hoped for result since the eigenvalues of the symmetric connectivity matrix are real and are independent of the numbering of the nodes. But this is known to fail for each of the three methods of assigning the diagonals listed above. It is true that the resulting eigenvalues "almost" distinguish the topologies except for a small percentage of networks which are called 'isospectral' meaning that the same set of eigenvalues represents two different topologies. But in the final analysis, although the connectivity matrix eigenvalue method distinguishes many of the topologies, it fails to distinguish a small percentage for n=5 nodes and higher n graphs.

Any arbitrary numbering of the nodes of a network allows an undirected network to be completely characterized by a connectivity (or adjacency) matrix C_{ij} which has the value '1' if nodes i and j are connected and '0' otherwise thus leading to a symmetric matrix. If the graph is directed with a connection pointing from i to j, then C_{ij} is set to 1 but with C_{ji} is set to '0' while diverse bandwidths of connectivity can be represented by any set of off-diagonal nonnegative reals. The setting of the diagonal can be to values of '1' if a node is considered connected to itself or '0' if it is not. Both assignments are relatively arbitrary and describe the same topology. Because of the arbitrary assignment of numbers to the nodes, there are n! different matrices (connected by the symmetric group on n symbols, S_n) that describe the same topology. Historically researchers had sought invariant measures (under S_n) with which to hopefully characterize the topology and solving such problems as graph isomerism (testing two C matrices to see if they describe the same topology). As the eigenvalues of C_{ij} are invariant under S_n then this has been a natural hope for network classification. Regrettably, there are networks that are topologically different as low as order 5 for which the eigenvalue spectra are the same (isospectral), whether the diagonals are set to value of '1' or '0'. Subsequent research has centered mainly on studies of the eigenvectors of C and powers of C. The author and associates have studied network topology of large structures using two different practical algorithms for cluster identification, one depending upon mutual generalized Shannon (Renyi) entropies and another model based upon ‘condensing’ clusters using a physical attractor model using forces between nodes which are uniformly distributed over an n-1 dimensional sphere. We have sought fast algorithms to monitor large networks in real time using Renyi entropies on clusters and subclusters as network metrics.

3.3 Lie Groups

This work will show that with a particular assignment of the (otherwise arbitrary) diagonal values of C_{ij} to be the negative of the sum of other values in each column, then C_{ij} is a member of a particular ‘Markov-type’ Lie algebra. Furthermore this result will hold even if the C_{ij} are generalized to any non-negative values and whether the graph is undirected (symmetric C_{ij}) or directed and thus nonsymmetrical. This result connects the static topology concepts of a network with the dynamical evolution of continuous (Lie) group theory. We will show that the C_{ij} represents the dynamical evolution of a conserved quantity (information, water, goods) on the network as the associated C_{ij} generates infinitesimal transformations that conserve the sum of the components of any vector upon which $M= \exp(\lambda_{ij} C_{ij})$ acts. Thus C_{ij} represents an infinitesimal transformation for a Lie group. It then follows that any non-negative values for C_{ij} and λ_{ij} are allowable and represent unequal flow rates such as would be the case for internet bandwidths or transportation flows. Then using the authors previous work on decomposition of the general linear group GL(n,R) into the Markov type group and an Abelian scale transformation group, we are able to generalize the diagonal to any real value which will be seen as a ‘creation’ or ‘annihilation’ rate at that node for the (otherwise) conserved quantity under the previous action. Generally then, our work takes any set of values of C_{ij} ,with non-negative off-diagonal values, and identifies the resulting transformation within GL(n,R) as a dynamic evolution of an associated Lie algebra. The meaning then becomes much more transparent for the interpretation of the eigenvalues and eigenvectors of C_{ij} similar to normal nodes for the system. For undirected topologies, the symmetric C_{ij} give eignenvalues representing the rates of exponential decrease to a state of higher entropy and equilibrium where all nodes have the same quantity of the conserved entity which we derive for a second order Renyi entropy.

We also investigate higher order products of the C_{ij} that specify nonrecurring paths back to a given node after k steps. Such non-recurring path generators provide useful new information on the lower order connectivity of the system related to clusters (sub-networks that are more highly connected than the surrounding network) and cliques (densely connected sub-networks with every node connected to every node). These structures are not linearly related to C_{ij} and provide eigenvalues and cluster values that go further to distinguish the topologies. Entropy values of these matrices provide additional invariant information and metrics for node ordering. The results can be collected as metrics for the topology and a deeper understanding of these metrics and the connection between Lie groups and network topologies. Generalization to any set of diagonal values is shown to be equivalent to adjoining the n-dimensional Abelian scaling group to the

Markov monoid thereby generating sources and sinks of the entity at each node independently that is otherwise conserved under the Markov monoid. As these new matrices each have differing eigenvalue spectra, but still characterize the same off-diagonal topology, they can be used to further classify networks.

It is necessary to understand some background material on Lie groups, algebras, and monoids including the general linear group and its subgroups. Specifically, our previous work explored a decomposition of the general linear group in n dimensions GL(n,R) into a Markov type Lie group M(n,R) and the Abelian group of coordinate scale transformations A(n,R) (or equivalently the Lie algebras that generate them). The n scale transformations are generated by the n by n matrices, L^{ii}, consisting of zeros except for a single diagonal position which has the value '1' as given by the matrix elements: $L^{ii}_{mn} = \delta^i_m \delta^i_n$. The Markov type Lie group was defined as those linear transformations which are continuously connected to the identity which preserve the sum of the components of a vector. The generators L^{ij} were defined in matrix form by the n^2-n different operators: $L^{ij}_{mn} = \delta^i_m\delta^j_n - \delta^j_m\delta^j_n$. The L^{ij} have the feature that the sum over any column is '0' and consequently, the Lie group M(n,R) that they generate, (with M = $\exp(\lambda_{ij} L^{ij})$), has elements where the sum over any column is '1' as is required for a Markov transformation. This previous work explored the Lie group generated by the Lie algebra defined by the L^{ij}. Then the general linear group can be easily seen to be generated by the two Lie algebras: the Markov type Lie algebra, L^{ij}, and the Lie algebra of scale transformations L^{ii}. The Markov type Lie algebra, is not simisimple, not nilpotent, and not Abelian. It possesses no Casimir type operators as all products of the algebra elements are contained in the algebra itself. M(2,R) is also the smallest nontrivial, non-Abelian algebra as it consists of only two elements with $[L^{12}, L^{21}] = L^{12} - L^{21}$. As the algebra preserves the sum of a vector's components, it leaves invariant the vector |1> = (1,1,…1) and constitutes motions on the hyperplane orthogonal to the vector |1> in n dimensions.

Special interest in this algebra arises from the restriction that all vectors with non-negative components are transformed into vectors with non-negative components for which the necessary and sufficient condition can be shown to be that the parameters λ_{ij} are non-negative in $\exp(\lambda_{ij} L^{ij})$. Such transformations constitute all Markov transformations which are continuously connected to the identity, but the non-negativity of the λ_{ij} result in the loss of the inverse transformations. It is known that the Markov transformations do not have an inverse and thus do not form a group, but this methodology allows one to utilize the power of Lie groups and algebras nevertheless to study continuous Markov transformations. The particular basis for the Markov Lie algebra chosen above, L^{ij}, performs the separation of allowable from non-allowable Markov transformations via the non-negativity of the λ_{ij} because the Lie Algebra basis was chosen to take a fraction of one component and give it to another. This type of transformation maintains and separates the acceptable transformations from those that can generate unphysical states (negative components). Consequently the Markov Lie group representations are very suitable objects for the manipulation of n-tuples that give the non-negative decomposition of unity required for probability theory.

3. 4 Markov Theory and Networks

Next we need to review the connection that the author and PI of this seedling grant has discovered between Markov theory and networks, namely that every network connection matrix, with the diagonals defined in a specific way, represents an element of a Markov Lie algebra monoid. In its abstract form, we can represent a network as a set of nodes that are numbered with the integers (1,2,…n) for identification. The connectivities among pairs of nodes are represented by lines which join nodes and their whole structure can be represented by the connectivity or adjacency matrix, L_{ij} which is defined to be equal to ‘1’ if nodes i and j are directly connected, and ‘0’ otherwise, where i and j range from 1 to n. The diagonal elements, L_{ii}, are traditionally set to ‘1’ if a node is considered ‘connected to itself’ or set to ‘0’ if it is not, thus defining the complete matrix with either ‘0’s or ‘1’s on the diagonal. We will here look only at undirected graphs for which the matrix is symmetric: $L_{ij} = L_{ji}$. Consideration of both the Markov Lie algebra generators and the connectivity matrix (with arbitrary diagonals) allows one to see that the connectivity matrix is a particular combination of the Lie algebra generators for the Markov type Lie group if the diagonals are set to be the negative of the off-diagonal terms in that column. This results in a matrix that has the sum of all elements in each column equal to zero. As the connectivity matrix is symmetric for an undirected network, the sum of elements in each row will also be zero. This method of setting the diagonal is in other contexts called the Lagrangian form of the connectivity matrix. In any of the three methods described above for the determination of the diagonal, it is obvious that each describes the connectivity in the same way, as connectivity is described by the off-diagonal elements. The central problem is that (even ignoring the diagonal terms) many different connectivity matrices describe the same ‘topology’ or connectivity among the nodes.   The root of this problem is that the numbering of the nodes, is arbitrary. The central problem then is to devise a mathematical technique to distinguish different networks or graphs and even more generally to classify all possible graphs of a given order (number of nodes) in a unique way and thus to eliminate the arbitrariness of the node number assignment but to not discard the essential ‘connectivity’ and thus to uniquely classify the topology itself. This is not only an unsolved problem but one that is known to be of extreme complexity and difficulty. The nodes must be numbered in order to form the connectivity matrix but all permutations of the numbering (and thus all resulting connectivity matrices) are topologically equivalent.

From the discussions on the Markov Lie algebra above, one immediately recognizes that if the diagonal elements of the connectivity matrix are taken to be the negative of the sum of the non-diagonal elements in that column, then the resulting connectivity matrix will be a generator of a continuous Markov transformation which in turn preserves the sum of components in a ‘vector’ upon which the transformation could act. In other words, the connectivity matrix is an element in the Markov Lie algebra. But since the network was defined originally as a static topology with no ‘vector space’ to be acted upon, such interpretation requires some reflection. The Markov transformation basically consists of values of ‘1’ for each valid connection between two nodes and ‘0’ otherwise and these can also represent transition probabilities per unit time. The implication is that if this matrix acts upon a vector that represents quantities at the nodes (water, information, energy, probability) then transfers will occur from all nodes at equal rates to the connecting nodes until equilibrium is reached. Thus the connectivity matrix, which initially represented a static topology, can now be interpreted as a time translation operator for the dynamical evolution of a vector of conserved substance that moves at equal rates among all connecting nodes. Obviously then, the eigenvectors of the matrix represent the equivalent of 'normal nodes' for the system. The eigenvectors are nodal combinations, of quantities at those nodes that exponentially decay at the rates given by the associated eigenvalue. Isospectral networks that represent different intrinsic topologies thus have identical decay rate spectra (similar to degenerate energy spectra for the Hamiltonian). Consequently, this analogy connects a static network topology, via the connectivity matrix, to dynamical flows of some conserved substance in the equivalent network. It furthermore provides an interpretation for alternative choices of the diagonal elements since any other choice other than that for the Markov algebra diagonal values implies an exponential growth or decay rate of the otherwise conserved substance at the respective node. Thus this analogy now also provides an understanding of the other choices for diagonal elements and can be used to appropriately model such dynamics. Although it is regrettable that these eigenvalues do not distinguish the topology, there might be other ways to break the isospectral degeneracy.

Now consider a connectivity matrix, C, with zero values placed on the diagonal. This matrix will give a transition from i to j when that value is '1'. By forming C^2, the terms give the number of ways that one can go from i to j through some intermediate node k. Specifically, $(C^2)_{ii}$, the diagonal elements, give a count of the number of ways that one can perform a transition from a node to another different node (since C has 0 for diagonal elements) and return to that node in exactly two steps. Now remove the diagonal values (replacing the elements with 0s), and use the removed diagonal values to form the first row of a matrix which we call the self-connectivity matrix, S. Next form $C^3 = C*C^2$ and place the new diagonal into the second row of the connectivity matrix and replace their values with 0 as before and continue this process. These diagonal elements of C^m count the number of ways that one can leave a node and return to that node in exactly m steps without passing through the initial node in the interim process. As such they 'feel out' the topological structure around each node by counting the number of distinct paths that leave and then return to each node without returning to that node during the transitions. As the maximal path to explore the entire topology requires n-1 steps out to the most distant node and n-1 back, then we need to perform this process 2(n-1) times (in order to acquire complete topological information) resulting in 2(n-1) rows to the self connectivity matrix. This process also results in 2(n-1) matrices whose ij elements count the number of paths from node i to node j and which we call the mutual connectivity matrices. The 2(n-1) mutual connectivity matrices, along with the original S matrix, can be converted into Markov algebra generators by placing the negative of the column sum in each diagonal position (replacing the 0s) thus yielding 2n-1 n x n matrices in this set. The process of removing the diagonal at each stage is not linear and thus the mutual connectivity matrices will have independent eigenvalues which are now to be computed for each matrix. These matrices generate continuous Markov transformations based upon generators that transition m steps at a time (leap out into the network) without returning to the original node in lowest order.

The n eigenvalues of each of these 2n-1 mutual-connectivity matrices will provide a more extensive classification of the topology independent of node ordering. Also the 2(n-1) different values of the sums over all nodes, for each order, of the self connectivity vectors, constitute another metric for the topology that is independent of the nodal ordering. Taken together this provides n(2n-1) + 2(n-1) = $2n^2$+n-2 independent measures of the topological structure that are independent of nodal numbering. The self connectivity matrix C provides extensive information on the nature of clusters around each node because if a node is highly connected to adjoining nodes at m steps, then the m, and subsequent self connectivity values for that node, will be very large. If on the other hand a node is isolated by only one path to other parts of the topology, then the values of that nodes' self connectivity will be minimal. In this last case one will have a minimum entropy and maximum information as defined as a sum over all nodes at each level m with a standard expectation of information (negative entropy) equation $I_m = \Sigma_j C_{mj} \mathrm{Log}_2 C_{mj}$). If the node is part of a clique (a sub-graph with every node connected to every other node) then the self connectivity will be a maximum (and thus will have a maximum entropy) and I_m will be maximal.

It is not claimed that these methodologies distinguish and classify all topologies but rather that they provide a way to methodically capture additional information on the topological structures over and above the eigenvalues and eigenvectors of the root connectivity matrix. We are currently using computer models to study these classifications. We also believe that the methods of using the various Markov Lie algebra elements derived from the connectivity matrix and the resulting dynamical time evolution of a conserved entity flowing on the network, gives useful insight into diverse practical network problems. Although the connectivity matrix studied above was for an undirected graph, the same analogies and models obtain with directed flows and also when the flows occur at different rates as represented by any set of real non-negative numbers in the off-diagonal positions of the connectivity matrix.

We have shown that the connectivity matrix for networks and graphs is part of a Markov Lie algebra (or monoid) and this provided insight into invariant measures of the topologies, clusters, and dynamics of networks and graphs. We now need to study how far these invariant measures can be used to uniquely classify topologies and sub-topologies in networks as well as study practical applications of the network dynamics. It is suggested that the Markov Lie group, associated Lie algebra, and generalized definition of information, can bring much deeper insight into an understanding of both classical and quantum uncertainty and their measurement both in practical problems and the theoretical foundations.

4 Results and Discussion

4.1 Patent Applications

The work of Gudkov and Johnson has resulted in a patent application ("System and Method for the Analysis and Classification of Networks and the Like") which is now pending involving entropy, mutual entropy, and cluster identification. The work of Johnson, Gudkov, and Nussinov had earlier resulted in a provisional patent application on the rapid identification of clusters in large networks; however, this provisional patent was not formally submitted. These clusters can be usefully used to track the associated entropy density. The attached technical documents and papers document the details of our work.

4.2 Conclusions

In conclusion, we seek methods of optimally defining $C_{ij}(t)$, and of selecting sub-networks (such as clusters, cliques, natural sub-nets, spectral disaggregation (as described above, and including the entire network itself) for each of which we calculate a series of generalized entropies and track these 'metrics' over time as macroscopic variables indicating network behavior. The hypothesis is that a judicious set of these choices will track the associated entropy changes over time such that anomalous values will be associated with network intrusions, attacks, or malfunctions.

The two primary tasks then become (A) the computation of the generalized entropies on networks that are mathematical simulations and the tracking of the metrics as the network changes in the connectivity and cluster structure over time. This work is critical because we can build networks of specific properties and study the behavior of the metrics in a controlled environment. The second task (B) is to study the behavior of the real internet traffic. While this has the advantage of being the 'real world', it has the disadvantage of being very difficult to actually measure. The measurement difficulties arise both from privacy issues and concerns as well as situations where the traffic simply cannot be gathered because it is among nodes all of which lie outside the domain of collection. Consequently, C_{ij} may only be known in certain blocks with others totally unknown.

4.3 Recommendations for Future Work

Implicit in future work is:

1. Optimal Entropy Metrics: Even with extensive progress already made, it is critical to reduce the number of metrics that can be reasonably investigated. We need to determine which of the generalized Renyi' entropies and which differences of which entropies.
2. Specifically we need a deeper understanding of entropy as is afforded by the links between topological structure, entropy and information theory, continuous group and

algebra theory, Markov theory, diffusion, and rates of change of the connection matrix.

3. As each Renyi entropy is defined on a vector of probabilities we need to know how to define this on a matrix (which is a collection of n vectors of probabilities for a probability distribution of n sectors).
4. We also need physical interpretation and rapid computational techniques to utilize for large matrices.
5. We need to understand the meaning of these entropies in terms of specific topological structures using mathematical analysis and prototyping experiments on randomly constructed matrices.
6. These prototype experiments must also be performed on changing topologies in order to determine the numerical ranges of specified changes in specific topologies.
7. We need to understand more about the range of real internet traffic flows as the subset of prototyped connection matrices. Specifically, what subcategories of possible scaling matrices (as these seem to limit the connections matrices describing real traffic).
8. It is necessary to both identify means of data collection and how to treat missing values and sectors of C that cannot be captured in real life.
9. Certainly one of the most important problems is how to identify subnets where the entropy metrics can be individually computed because the entropy deviations are very small on very huge networks until it is too late and the attack or aberrant process has dominated a large portions.
10. Given this, then upon real subnets what is the normal range for the chosen entropy metrics being tracked on the subnets.
11. It would be important to investigate the extent to which one can identify a particular worm, virus, or malicious by the metrics profile.
12. In conclusion: our recommendations center on increasing the mathematical understanding to identify optimal entropy and gaining an intuitive guide, increasing the speed of the computation, gaining knowledge, as specified above, from the prototype experimentation, and finally proving the primary hypotheses, on real networks, that entropy metrics give a general, fast, intuitive, hierarchical, method for providing at least an order of magnitude improvement in the detection of malevolent processes on computer networks especially new types of attacks.

Appendix A

Approaches to Network Classification

Vladimir Gudkov* and Joseph E. Johnson†

Department of Physics and Astronomy
University of South Carolina
Columbia, SC 29208

Shmuel Nussinov‡

Department of Physics and Astronomy
University of South Carolina
Columbia, SC 29208 §

Zohar Nussinov¶

Theoretical Division,
Los Alamos National Laboratory
Los Alamos, NM 87545

(Dated: January 27, 2005)

Abstract

We introduce a novel approach to description of networks/graphs. It is based on an analogue physical model which is dynamically evolved. This evolution, which is numerically simulates, depends on the connectivity matrix and readily brings out many qualitative features of the graph.

PACS numbers:

*`gudkov@sc.edu`

†`jjohnson@sc.edu`

‡`nussinov@ccsg.tau.ac.il`

§on sabbatical leave from Tel-Aviv University, School of Physics and Astronomy, Tel-Aviv, Israel

¶`zohar@viking.lanl.gov`

I. INTRODUCTION

A graph or network consists of n vertices/nodes V_i with edges (communication lines) connecting them. It can be described by an $n \times n$ connectivity matrix C (refereed to in graph theory as the adjacency matrix) where

$$C_{ij} = C_{ji} = \text{number of edges connecting } V_i \text{ and } V_j.$$

Even when we allow C_{ij} to be only 0 or 1 - for (dis)connected V_iV_j, the number of C matrices $2^{n(n-1)/2}$ is huge already for moderate n.

If two matrices differ only by the labelling of the vertices - i.e. by a similarity transformation $C' = U^{-1}CU$ with U ($U^{-1} = U^{\dagger}$) effecting the permutation of rows (columns) of C - then C and C' represent the same graph.

Since there are n! such permutations the problem of deciding whether the two connectivity matrices correspond to the same graph though not NP complete [1] is believed to be of a high degree of difficulty. It is equally hard to find intrinsic relabelling invariant features, of graphs which characterize *all* graphs. Even if not achieving this goal, such intrinsic features may be most valuable. Thus the characteristic polynomial or eigen-value $(\lambda_1 \ldots \lambda_n)$ of the connectivity matrix encode many important graph theoretic features[2].

For most applications a complete characterization of graphs/networks is redundant. We are often interested in the "Big picture" or gross features. These include the answers to the following general questions about the graph/network:

$\mathbf{Q_1}$: "Are there some groups of vertices which are relatively strongly interconnected and more weakly connected to the rest of the "external vertices" ? "

We will refer to these groups as "clusters in graph." Clearly these differ from the graph theoretic "cliques" defined by requiring that each vertex in the clique be connected to all other vertices in the clique with no reference to the extent of external connections.

$\mathbf{Q_2}$: "Are there groups of vertices which are "distant" from each other in the sense that there are no (or few) "short paths" connecting them?" ("Short paths" are those with a small number of consecutive links.)

Ideally we would like to view a complex graph as a smaller set of ($k \ll n$) of "super vertices" each having a specific internal structure. By connecting to other super vertices, these form a "super graph" at a higher level.

The shear number of graphs defied such a goal when *all* graphs are considered. We believe however that actual communication, social, commercial, political etc networks are essentially *not* random.

The very history of their, often gradual, formation can result in a hierarchial clustering. There is often a further tendency to enhance clustering. If V_i and V_j are both strongly connected to V_k then V_i and V_j also frequently develop a direct connection.

Physical constraints such as the three dimensional space we live in and the essentially two dimensional surface of the earth or boards of printed circuits also play a crucial role along with the need to economize on the total length of communication lines used.

All the above tends to make "clusters in graphs" with relatively loose connections between them more likely so that two questions Q_1, Q_2 above can be answered in the affirmative.

The following analogy may be instructive. An outstanding, problem in post genomic biology is to predict the folding of proteins given their known amino acid sequence. While natural "native" proteins almost instantaneously fold into their functional three dimensional form, artificially constructed, random, sequences do not. It is believed that specific smooth "energy landscapes"[3, 4] help guide the system to its correct folded form - in nature and in simulations. This is reminiscent of the present problem where methods geared to specific "Real Life" networks with a presumed tendency for clustering are advantageous.

How can we efficiently search for such patterns?

We can ask for the number of paths in the graph of length s connecting a vertex V_i to itself or V_i to V_j. By "feeling out" larger and larger region (as s increases) we can tell if V_i belongs inside a cluster and if V_i and V_j are distant in the sense described above. We will elaborate on a simple approach for achieving this in Section II below.

Bringing vertices in a "cluster" into close spatial proximity can help in identifying these clusters. This can be achieved in a dynamical approach in which we model the vertices V_i by moveable point masses at $\vec{r}_i(t)$. Attractive "forces" are postulated between any pair of points which are connected in the original graph. Possible implementations of this general approach are discussed in Section III, constituting the main novel part of this paper.

II. THE NUMBER OF RETURNING PATHS AS A TEST FOR "CLUSTERING IN GRAPHS"

Imagine an actual physical model of the network where each edge is replaced by a 1Ω resistor. The electrical resistance between two nodes (or between two groups of vertices which are separately shorted) nicely models the "distance" between these nodes (or the two groups) as defined in Q_2 above. The laws of adding resistances in series and in parallel imply that the resistance, like the "distance", increases the longer the paths on the graph connecting the two nodes are, and also decreases with the number of such connecting paths. Instead of using this analog computation we can, by using powers of the connectivity matrix C, trace out the evolution in s steps of messages sent from each node to all its neighbors. In fact the i, j elements of C^s; namely $(C^s)_{ij}$ equals the number of paths comprised of s connected edges which start in V_i and terminate in V_j. In particular $(C^s)_{ii}$ is the number of paths returning to V_i in s steps.

When raised to a high power (smaller than n) C, like any symmetric real matrix tends in general to simplify considerably [9]. Let $\lambda_1 \ldots \lambda_n$ be the n real eigenvalues of C in descending order and $\vec{V}_1 \ldots \vec{V}_n$ the corresponding orthonormal n eigenvectors. The columns $\vec{C}^s$ of C^s become all proportional to $\vec{V}_1$ with a factor representing the projection of $\vec{V}_1$ on the i-th column of C:

$$(\vec{C}^s)_i \propto (\vec{C}_i \cdot \vec{V}_1)\vec{V}_1 \tag{1}$$

and likewise for the rows. Upon further multiplication by C, C^s gets then multiplied by λ_i.

For the special case when all vertices in C have the same valency v (i.e. each is connected to v others) $\lambda_i = v$ and

$$V_i^+ = \frac{1}{\sqrt{n}}(1, 1, \ldots 1). \tag{2}$$

While we seek some dilution of information such trivialization should be avoided. Useful information can be obtained by looking at $(C^s)_{ij}$ at moderate values of s. If i belongs in a rich heavily connected, "cluster in graph" then with an valency in cluster v_{cl} the initial rise of $(C^s)_{ij}$:

$$(C^s)_{ii} \sim (v_{cl})^s \qquad for\ i \in cluster \tag{3}$$

is higher than the initial rise of the same quantity when V_i is a generic vertex located in a region of average ($\overline{v}$) valency so that:

$$(C^s)_{ii} \sim (\overline{v})^s \qquad for\ i \notin cluster \tag{4}$$

with $\overline{v} \leq v_{cl}$.

To partially avoid the degeneration at high s, and gain more information from C^s for large s, we tried adopting the following strategy. Instead of C^2 we use

$$\tilde{C}^2 = C^2 - diag\ C^2 \tag{5}$$

Since the diagonal of C^2 counts all paths which come back to their origins in two steps, these paths are omitted in $\tilde{C}^2$. Going one more step we consider $\tilde{C}^2 \cdot C$. By subtracting again its diagonal elements and defining

$$\tilde{C}^3 = \tilde{C}^2 \cdot C - diag\ \tilde{C}^2 \cdot C \tag{6}$$

we omit all paths which retrace in three steps, etc. In general we define

$$\tilde{C}^{s+1} = \tilde{C}^s \cdot C - diag\ \tilde{C}^s \cdot C \tag{7}$$

And $(\tilde{C}^{s+1})_{ij}$ is the number of paths from i to j of length $s+1$ which have not revisited at any prior stage the initial V_i[**?**], and $(\tilde{C}^s \cdot C)_{ii}$ is the number of $i \rightarrow i$ such paths.

While the latter number increases more slowly than $(C^s)_{ii}$, it still "runs-away" as $s \rightarrow \infty$, so that we need to "re-normalize" $\tilde{C}^s$ at each stage to have each $(\vec{\tilde{C}^s})_i$ column vector be of unit length. A plot of $(\tilde{C}^{s-k} \cdot C)_{ii}$ as a function of s could ideally help "map out" other clusters in the graph. After staying for some steps in the putative initial cluster C_1 in which i resided we will wander off into a generic part of the graph. There the slower growth rate (5) will take over. If we can reach in d_{i2} steps a second rich cluster C_2 we could after such number of steps start having again a fast growth rate (3).

However the graph "between the clusters" is still a network. This causes diffusive migration between two clusters with no sharp arrival times. Also for appreciable s several clusters may be reached at the same or similar number of steps. These features tend to smooth out the changes of $(\tilde{C}^s)_{ii}$.

III. DYNAMICAL EVOLUTION HIGHLIGHTING NETWORK STRUCTURE.

A basic difficulty in discerning intrinsic graph / network structure is that the connectivity matrix depends on the labelling of the vertices. The following example clearly illustrates this. Let us assume a large subset of vertices in our graph indeed divide naturally into fairly well-defined clusters C_1 with n_1 vertices, C_2 with n_2 vertices etc up to C_k. If we label our vertices in such a way that all vertices belonging in any one cluster are contiguous, the connectivity matrix will be "Almost Block Diagonal".

This is depicted in Fig.(1) where the non zero (unit) entrees of the C matrix are represented by a dot at the coordinate (ia, ja) and the 0's by having an empty $a \times a$ square at the point (ia, ja). The $n_1 \times n_1$, $n_2 \times n_2$ sub matrices along the diagonal will then be connectivity matrices for the first, second, etc cluster. By assumption these matrices have a relatively high proportion of non-vanishing elements. The corresponding darker squares can thus be visually discerned relative to the background of the lighter more, sparsely populated, remaining parts of the original C matrix.

This nice feature completely disappears after massive relabelling, i.e. massive joint reshufflings of columns and rows in the matrix C (Fig.(2)). The initial cluster seems to have disolved and the whole matrix will then have a roughly constant average density of dots i.e. unit entries looking uniformly grey. Our goal is essentially to reconstruct the original, convenient "Almost Block Diagonal" form which exhibits the clusters. Its difficulty is exacerbated by the fact that the block diagonalization is only approximate and there are many non-vanishing entries outside the blocks. Also we do not know a priori which size blocks and how many blocks exist.

The representation of a graph by drawing it in two dimensions also introduces undesired arbitartrariness reflected in the choice of coordinates (x_i, y_j) of the points representing the various vertices. Two different drawings of the same graph may appear completely different and unrelated.

Such arbitrariness is particularly harmful when we try to reconstruct the clasters by introducing attractive forces between any pair of points representing a pair of connected vertices. The subsequent motion of the points does depend on their arbitrary initial placement.

To place the n vertices in a completely symmetric and unbiased manner we need to go to $n - 1$ dimensions. The vertices (or the n physical point masses modelling them in our approach) can be then put at the n vertices of a symmetric simplex inscribed inside the unit sphere in $n - 1$ dimensions. Specific coordinates of the n vertices can be constructed in a simple inductive process indicated in Appendix A. All vertices are equidistant from the

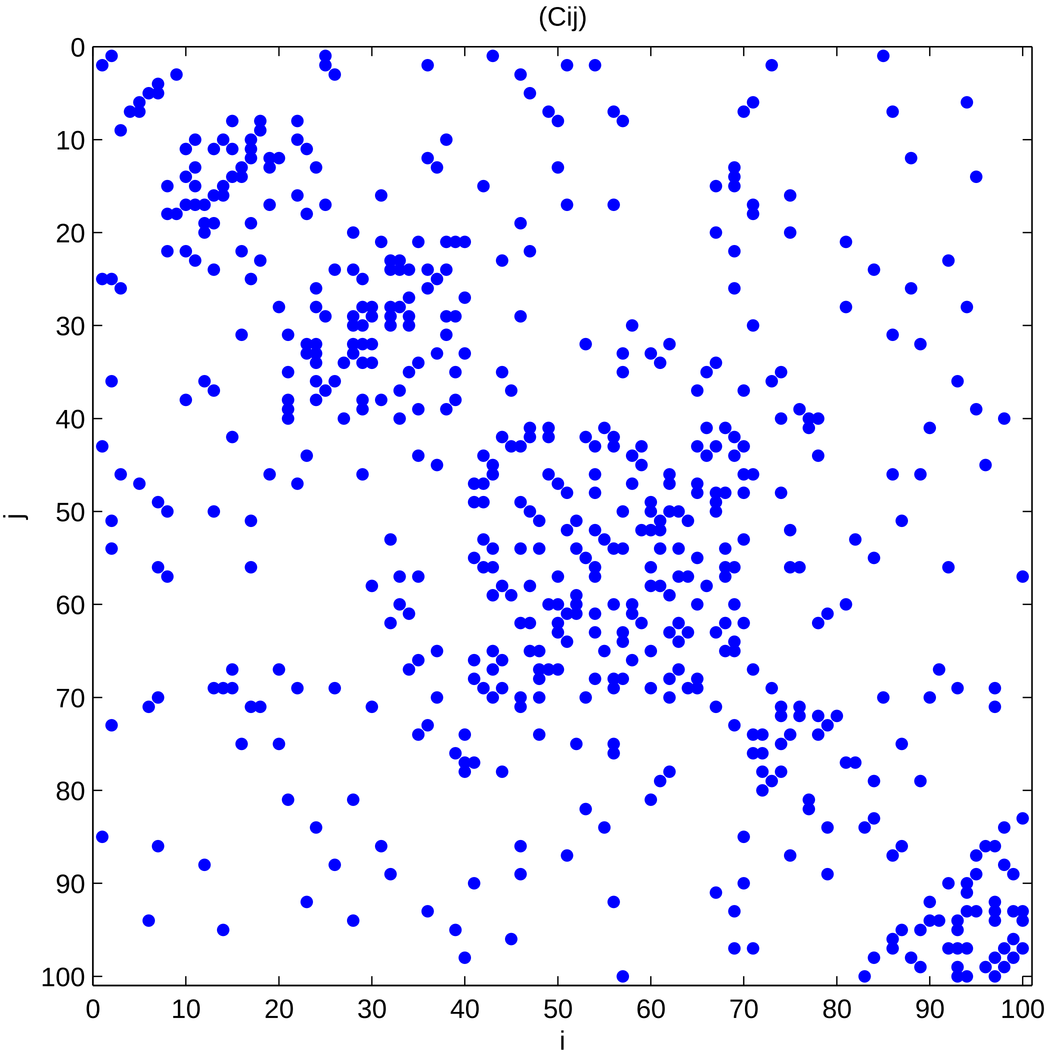


FIG. 1: Connectivity matrix with the average cluster valency 20% and inter cluster connectivity valency 3%.

origin; and specifically we chose:

$$\vec{r}_i^{\,2} = 1 \tag{8}$$

using this,$(\sum \vec{r}_i)^2 = 0$ and the equality - due to symmetry - of all $\vec{r}_i \cdot \vec{r}_j$ for any $i \neq j$ readily implies:

$$\vec{r}_i \cdot \vec{r}_j = -\frac{1}{n-1} \qquad all \quad i \neq j \quad i, j = 1 \dots n. \tag{9}$$

The distance between any pair of vertices of the simplex i.e. between any pair of the representative points at the outset of our proposed dynamical simulation is therefore:

$$|\vec{r}_i - \vec{r}_j| = \sqrt{\frac{2n}{n-1}} \qquad all \quad i \neq j. \tag{10}$$

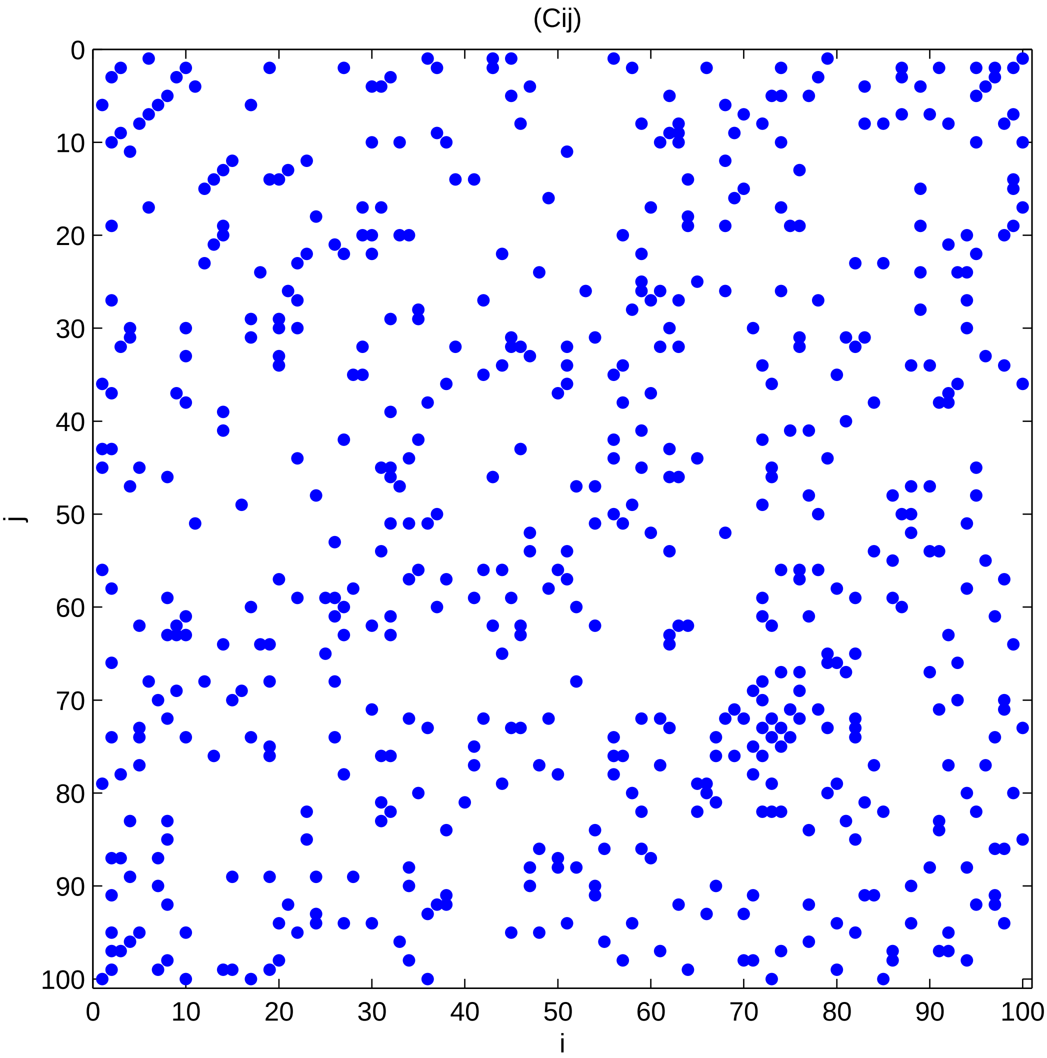


FIG. 2: Randomly reshuffled connectivity matrix C.

We next endow our system with some dynamics[7]. We introduce a fictitious attractive force between points corresponding to vertices which are connected in the initial graph of interest. Thus if $C_{ij} \neq 0$ we postulate

$$\vec{F}_{ij}(\vec{r}_i, \vec{r}_j) = \zeta_{ij} f(|\vec{r}_i - \vec{r}_j|) \frac{(\vec{r}_i - \vec{r}_j)}{|\vec{r}_i - \vec{r}_j|}. \tag{11}$$

To be the force attracting the point mass i to the point mass j, in the direction of $\vec{r}_i - \vec{r}_j$. To retain the initial symmetry and avoid any biasing we take the same force law $f(r)$ for all pairs. The specific shape of $f(r)$ can be tuned to optimize the gradual clustering. In general $f(r)$ falls with distance is now inoperative In the following we uses constant forces ($f = const$).

The only way information about the specific graph of interest is communicated to our dynamical n body system is via the overall strengths of the forces ζ_{ij}. It vanishes if $C_{ij} = 0$. For the generalizations considered later and to mimic real networks we allow any $\zeta_{ij} \equiv C_{ij} > 1$ so that it counts the number and "quality" of connections between V_i and V_j.

We next let our point move according to standard newtonian dynamics:

$$m_i \frac{d^2\vec{r}_i}{dt^2} = \vec{F}_i = \sum_j \vec{F}_{ij}. \tag{12}$$

To avoid "overshoots" and oscillations we add damping via viscous frictional forces:

$$m_i \frac{d^2\vec{r}_i}{dt^2} + \mu_i \frac{d\vec{r}_i}{dt} = \vec{F}_i. \tag{13}$$

Finally we adopt the extreme $\mu_i \gg m_i$ so as to neglect inertial effects and have first order "Aristotelian Dynamics":

$$\mu_i \frac{d\vec{r}_i}{dt} = \vec{F}_i. \tag{14}$$

The latter is readily discretized for time increments δ:

$$\vec{r}_i(t+\delta) = \vec{r}_i(t) + \frac{\delta}{\mu_i}\vec{F}_i(\vec{r}_i(t)) \quad i = 1 \ldots n, \quad l \neq i \quad l = 1 \ldots n. \tag{15}$$

To preserve the initial symmetry we take all initial mass (and separately all initial viscosities) to be equal $\mu_i = \mu$, $m_i = m$. Different masses (and / or viscosities) will arise at later stages when we treat super graphs with heavy vertices representing initial clusters.

The attractive central forces can be derived from a pair wise potential i.e.:

$$f(r) = -\frac{d}{dr}U(r). \tag{16}$$

And the overall potential energy is then:

$$U(\vec{r}_1 \ldots \vec{r}_n) = \sum_{i>j} \zeta_{ij} U(|\vec{r}_i - \vec{r}_j|). \tag{17}$$

The constant forces alluded to above arise when we have linear pair-wise potential or fixed tension wires connecting the points[**?**]. The possible equilibrium "fixed points"of our dynamical system namely those for which

$$\frac{d\vec{r}_i}{dt} = \vec{F}_i = 0 \tag{18}$$

for all i are then stationary points of $U(\vec{r}_1 \ldots \vec{r}_n)$.

With only attractive forces or potentials present our n point system eventually collapses towards the origin. This is readily seen as the scaling

$$\vec{r}_i \to \lambda \vec{r}_i \tag{19}$$

with $\lambda < 1$ will obviously decrease the $U(\vec{r}_1 \dots \vec{r}_n)$ of equation (17) for any set of $\vec{r}_i$.

A collapse of all n points happening before the vertices belonging to "clusters in the graph" have separately concentrated in different regions defeats our goal of identifying the latter clusters.

To avoid the radial collapse we constrain $\vec{r}_i(t)$, to be at all times on the unit sphere:

$$|\vec{r}_i(t)| = constant = 1 \quad all \quad t \geq 0. \tag{20}$$

To incorporate this we supplement eq.(15) by a length renormalization:

$$\vec{r}_i(t+\delta) \to \frac{\vec{r}_i(t+\delta)}{|\vec{r}_i(t+\delta)|} \tag{21}$$

to be performed following the operation (15) at each step of our evolution. The constraint (20) amounts to introducing normal (radial) reaction forces which cancel the radial components of any of the forces $\vec{F}_i$, leaving us with only the tangential parts:

$$\vec{F}_i^T \equiv \vec{F}_i - (\vec{F}_i \cdot \vec{r}_i)\vec{r}_i. \tag{22}$$

While the above avoids the radial collapse, the residual tangential forces can still initiate a collapse at some point on the unit sphere.

The basic conjecture we make is the following: "After a sufficiently long time T (or sufficiently many steps $s = T/\delta$) has elapsed so that any point moved on average an appreciable distance away from its initial location $|\vec{r}_i(T) - \vec{r}_i(0)| \geq a \approx 1$ geometrical clusters of points tend to form. The points in each geometrical cluster correspond, to a good approximation, to the original vertices in a "cluster of the graph" which these points represent."

In the following we motivate this conjecture, and test it numerically.

We recall the definition of a cluster in the graph as a subset C_l of n_l vertices with a higher number of connections between them than the average number of connections with "external" vertices, which are not in the cluster. At $t = 0$, the points representing any subset of p vertices out of the n vertices in the graph reside at the p vertices of a $(p-1)$

dimensional symmetric simplex. All together there are $\begin{pmatrix} n \\ p \end{pmatrix}$ such "faces" of our original $n-1$ dimensional simplex.

To most clearly illustrate our point let us assume an "ideal graph cluster" with $p = n_l$ vertices so that in first approximation we neglect forces attracting members of the cluster (more precisely point masses representing vertices in the cluster) to "outside" points. Had we also omitted the constraint (20) then the forces acting between the n_l points of the cluster C_l would initially and hence at all subsequent times, be restricted to the corresponding n_l-1 dimensional face. Repeating the argument made originally for the full set of n vertices, a collapse of these n_l points into some point inside the n_l simplex (i.e. on the n_l-1 dimensional face) is guaranteed. With the constraint (20) enforced, the set of n_l points will still collapse but now not to a point on the n_l-1 simplex but to a point on the "spherical n_l-1 simplex" which is the projection of the n_l-1 simplex on the unit sphere. The point of common clustering need not be at the geometrical center of this spherical n_l-1 dimensional face. However unless the cluster in question is very asymmetric in its internal connections, it may not be too far from it.

Let us next turn on the remaining fewer forces pulling members of the cluster due to external vertices, i.e. towards points initially residing outside this face. Such pulls may shift the location of the clustering point away from the n_l-1 dimensional spherical "face". It is unlikely that it will disrupt completely the clustering of the vertices $V_i \in C_i$ belonging in the cluster.

We believe that the tendency to cluster will persist even in the more general case when the clusters are not so sharply defined.

Let us focus on one particular vertex V_i located at $t=0$ at $\vec{r}_i$, one of the n vertices of the $n-1$ simplex vertices. Among all the $\begin{pmatrix} n \\ n_l \end{pmatrix}$ subsets of n_l vertices, i.e. n_l-1 dimensional faces, a subset of $\begin{pmatrix} n-1 \\ n_l-1 \end{pmatrix}$ shares the specific V_i. Stated differently, $\begin{pmatrix} n-1 \\ n_l-1 \end{pmatrix}$ different n_l-1 dimensional faces do intersect at the V_i considered i.e. $n-1$ edges, $\frac{(n-1)(n-2)}{2}$ triangles, $\frac{(n-1)(n-2)(n-3)}{3!}$ tetrahedrals and so on. Furthermore each of the triangles includes two of the $n-1$ edges impinging at V_i, every tetrahedra contains three of these edges, etc.

Let us next assume that among all such $\begin{pmatrix} n-1 \\ n_l-1 \end{pmatrix}$ simplexes there is a particular one which we denote by S_l so that the point in question $\vec{r}_i$, has a maximal number of forces acting in its direction (as compared with the number of forces acting on the direction of any one of the other simplexes). This is the reflection in our dynamical model the fact that the vertex V_i belongs in a cluster C_l i.e. has more connections to $V_j \in C_l$ than to vertices in any other subset of n_l vertices.

It is obvious that in the initially symmetric situation the point $\vec{r}_i$ will then start moving in the direction of that specific $n_l - 1$ dimensional face since the force in its direction will be maximal. The motion will not be *exactly* in this hyperplane as V_i may have some external connections and consequently there will be forces on the mass point $\vec{r}_i$ in other directions. However since the largest force component is along this direction so will be also the largest initial displacement $\delta_1(\vec{r}_i) = \vec{r}_i(\delta) - \vec{r}_i(0) \;\propto\; \vec{F}_i$. This motion will then be the first small step towards the formation of the physical cluster of the points representing C_l.

At $t = 0$ all vertices start moving. If all (or most) of the n_l vertices on the simplex (face) in question share this same feature of V_i then all (or most) of the n_l points will tend to migrate away from the initial n_l vertices of the simplex in question and move toward its interior. Once the representative points start to cluster on or near the corresponding $n_l - 1$ dimensional spherical face the non-linear aspects of the many-body dynamical evolution come into play. These will tend to enhance and accelerate the clustering.

As the group of points start to come closer together the average distances $|\vec{r}_i - \vec{r}_{i'}|$ with $(V_i, V_{i\prime}) \in C_l$ decrease. If the attractive forces between them become stronger this accelerates the clustering of the points which started to cluster. Since in our present application we used constant, distance independent, $f(r)$, this is not operative here. We still have the very important additional effect namely the more coherent pull on “straggling vertices” by the other vertices belonging to a strong cluster. (“Straggling, or straying vertices” are those which due to “accidental” connections to some different group of vertices start moving in a different direction, than that of the face in question.)

The initial forces acting on any vertex have an angle of 60^o between any pair. However because of our constraint of staying on the sphere we need to consider only the projection on the $n - 2$ hyperplane tangent to the sphere at the vertex V_i, say, in question. After this projection the $n - 1$ edges emanating from V_i span the $n - 2$ dimensional hyperplane just in the same symmetric manner as the n unit vectors $\vec{r}_i$ span the original $n-1$ dimensional simplex. Hence at eq.(9) the angle between members of any pair of the effective tangential forces is

$$\cos\left[\theta_{ij[projected]}\right] = -\frac{1}{n-2}. \tag{23}$$

Thus if V_i was connected to *all* the remaining $n-1$ vertices in the original graph the sum of all the (tangential!) forces acting on it would vanish. In reality the valency of V_i, v_i = total number of vertices directly connected to it is, much smaller than $n-1$. Still the almost orthogonal v_i forces acting on it will thus tend to add in *quadrature.* The same a-fortiori holds for the v_{iC_l} forces directed to the face representing the cluster C_l. (v_{iC_l} is a partial i-C_l valency, namely the number of vertices in C_l connected to V_i).

The initial force component along the $n_l - 1$ dimensional face is then:

$$\vec{F}_{i\{i\in C_l\}}(t=0) = \sum_{j\in C_l} \vec{F}_{ij}(t=0) \ \propto \ [v_{iC_l}]^{1/2} \tag{24}$$

Assume however that after some time most points corresponding to the putative cluster, and, in particular, the v_{iC_l} points in the cluster C_l connected to V_i, have already bunched together on the surface of the sphere. The various forces exerted by these v_{iC_l} points on V_i will now be almost *parallel* and instead of (24) we will have

$$\vec{F}_{i\{i\in C_l\}}(t>t_0) = \sum_{j\in C_l} \vec{F}_{ij}(t>t_0) \ \propto \ v_{iC_l}. \tag{25}$$

Hence the resulting force will be considerably enhanced if $v_{iC_l} >> 1$!

If the vertices in the original graph had on average small overall valency then v_{iC_l} could happen to be small - say $O(2-3)$. The $\sqrt{v_{iC_l}}$ enhancement of (25) relative to (24) would then be minimal. Also v_{iC_l} could be smaller than the number of connections that V_i happens to have with points in some random face with $n_{c'1} = n_l - 1$ dimensions. The vertex V_i will then "wander off" at $t=0$ in the direction of this face rather than that of the "correct" face corresponding to the cluster C_l. We can avoid such situations and enhance the coherence effect discussed above by replacing the original connectivity matrix by an appropriate power (C^s or $\tilde{C}^s$ defined in Sec. II above) where the overall valencies (and in particular valencies pertaining to cluster) are (particularly) strongly enhanced.

Note that an "error"due to an initial wandering off of V_i in the direction of some random face which corresponds to no cluster in the graph, is corrected by the very clustering which is assumed to occur. The other points in the "random" face will, by assumption, tend to migrate out of this face into other faces where these points can more efficiently cluster (physically). Finding no nearby points on the wrong face the "straying" vertex V_i in question is likely to be pulled back into the original cluster C_l (or to another cluster which formed in the meantime and to which V_i is more strongly connected).

Thus our dynamical evolution process is not just motion of n points on the unit $n-1$ dimensional sphere, rather we can view it as a competition between the putative different (physical!) clusters for additional members (points). In this ongoing "tug of war" clusters with stronger internal connectivity are likely to "win over" farther members and form first.

Once the points corresponding to a cluster in the graph have "bunched" close together they become effectively one dynamical unit - a "supervertex". Not only will all the points pull coherently external points but also the converse naturally holds: the clustered points will tend to respond coherently as one dynamical unit to an external force. Thus assume that we try to "pull away" one member point. Due to its many connections to the other members of the cluster the point in question will strongly pull on those connected to it. The latter points in turn will pull on further points in the cluster etc and eventually the whole cluster will move in response to the external force.

The actual emergence of the physical clusters can be readily ascertained. Once $|\vec{r}_i - \vec{r}_j|$ is smaller than a prescribed small number ϵ, and further more persist in staying that closed for some number s_p of steps the pair of points are "merged" into one point, at $(\vec{r}_i + \vec{r}_j)/2$. (Actually we need at this point to project again $(\vec{r}_i + \vec{r}_j)/2$ onto the sphere.) In further evolution the force acting on the merger point is the sum total of all the forces acting on $\vec{r}_i$ and $\vec{r}_j$. Also the resulting point should be endowed with twice the viscosity and twice the inertia $\mu_{i\cup j} = \mu_i + \mu_j$ and / or $m_{i\cup j} = m_i + m_j$. This new, doubled up, point represents a new graph derived from the original by identifying V_i and V_j. It has $n-1$ vertices and its connectivity matrix has the same elements $C_{gg'}$ when both gg' differ from either i or j. The new (compound) vertex ($V_{i\cup j}$) is now connected to all vertices which were connected to either i or j.

We can keep on merging, using at each step the center of mass of the points in question

$$\vec{r}_{cm(i,j)} \equiv \frac{m_i\vec{r}_i + m_j\vec{r}_j}{m_i + m_j} \tag{26}$$

as the merge point. We add up the masses and viscosities of the merged points and keep the connections to all vertices/points presently existing.

Ideally this process would yield, after a reasonable number of steps s, k "supervertices" corresponding to the k blocks $\{n_{l_1} \times n_{l_1},\ n_{l_2} \times n_{l_2}, \ldots\ n_{l_k} \times n_{l_k}\}$ in the properly ordered original C matrix of Fig. 1. The off-diagonal element ll' will be here the total number of non vanishing unit entries in the original matrix C in the $n_l \times n_{l'}$ rectangle at the "intersection" of the C_l, and $C_{l'}$ blocks. We could now repeat a similar dynamical procedure for the k "supervertices" This is in fact what the above algorithm is doing anyway in a relatively smooth and continuous manner. Indeed, even prior to the actual act of merging, the set of points in a cluster act coherently as one unit.

Instead of merging pairs of close by points, we can identify various physical clusters with some minimal number of points and merge those as above.

The dynamical evolution described here forms clusters of all sizes: small ones with few members, larger ones which may include all or parts of smaller clusters and the one big supercluster containing all n vertices.

In structure formation in three dimensions, creation of small clusters requires the particles forming the cluster to travel for shorter distances than in the case of bigger clusters. Due to the peculiar geometry of the initial symmetric in $n-1$ dimensional symplex described in appendix A, this intuition does *not* carry over. Formation of any cluster requires roughly the *same* distance to be covered regardless of the size of the cluster. Hence we are not guaranteed by essentially kinematic reasons that the smaller clusters will form first - en route to the bigger clusters [**?**] which is the desired scenario for our purposes here.

How should we tune the force $f(r)$ in order to help achieving such a scenario? If $f(r) \simeq c/r^{\alpha}$ with large α , small differences in the distances will have large effect, (note that for the gravitational force in $n-1$ dimensions, $\alpha = n-2$). Too strong a rise for $r < r_{initial}$ and fall for $r > r_{initial}$ may however lead to accidental clustering of some small groups. In particular it may diminish the effect of the corrective mechanism described above of via the coherent pull of the elements of the cluster on straying elements.

The following procedure prevents complete clustering but allows formation of clusters with higher than average internal connectivity. We introduce in addition to the above attractive force between vertices V_i and V_j with $C_{ij} \neq 0$, repulsive forces when V_i and V_j are *not* connected:

$$G_{ij} = g(\vec{r}_{ij})\frac{(\vec{r}_i - \vec{r}_j)}{|\vec{r}_i + \vec{r}_j|} \equiv F_{ij}(r) \quad for\ Cij = 0. \tag{27}$$

Again this can be derived from a repulsive $W(r)$ potential.

Since in general we have many more unconnected vertices in a graph with large n, the repulsion can be weak relative to the attraction. Let us assume that the average valency is v. If all n points would physically cluster we have $O(n^2/2)$ repulsive interactions $W(a)$, with a the size of the clustering region, and $O(nv/2)$ attractive interacter $V(a)$. Thus it suffices to have

$$W(a) \geq \frac{v}{n}V(a) \tag{28}$$

in order to prevent forming complete clustering into one big supercluster. (The constraint $|\vec{r}_i(t)| = 1$ is still necessary to prevent vertices from being pushed to infinity!)

Assume that a putative new member is trying to join a cluster C_l, in which its valency $v_{i\{C_1\}}$ is higher than the average. To facilitate joining, we need to satisfy the following condition:

$$W(a) \leq \frac{v_{i\{C_1\}}}{n_1}V(a). \tag{29}$$

Since $v_{i\{C_1\}} \geq v$, and further $n_l << n$, we have a sizeable range of $W(a)/V(a)$ for which smaller clusters but not very large ones can first form. By gradually phasing out the repulsive forces once the smaller clusters have formed, we can proceed to forming bigger clusters etc.

We note that repulsive forces tend to move to antipodal points on the sphere groups of points which are "distant" from each other in the graph theoretic sense of question 2 in the introduction.

IV. SPECIFIC APPLICATIONS

We applied the above approach to the problem of cluster identification in the 100-nodes network represented by the connectivity matrix C of Fig.(1). This matrix consists of seven clusters with randomly created internal connections with valency 20%. These clusters have been randomly interconnected via a background valency of 3%. To simulate a real-life situation of networks with unknown structure (topology) we randomly permute the rows and columns of the matrix C obtaining the reshuffled matrix C' shown in the Fig.(2). Next we apply our algorithm for clusters reconstruction using a combination of attractive and repulsive forces in $n-1=99$ dimensional space. The vertices of the 100-simplex were allowed to move under the influence of the forces on the 98-dimensional hyper-sphere in 99-dimensions. After a number of steps (about of 100) we analyzed the mutual distances between the vertices of the simplex and grouped neighbors which are close to each other (within a relative distance of order 0.1) into separate clusters. The new connectivity matrix is shown in Fig.(3). We see the seven "big" clusters of the matrix C on a background of few small additional clusters due to the random (but still rather high) cluster inter connections. The procedure identifies not only the cluster structure of networks but numerates and tabulates all the nodes in each cluster. We father note that the distances in Fig. (3) between the different clusters do - unlike in the original Fig. (1) - reflect the actual "graph theory"

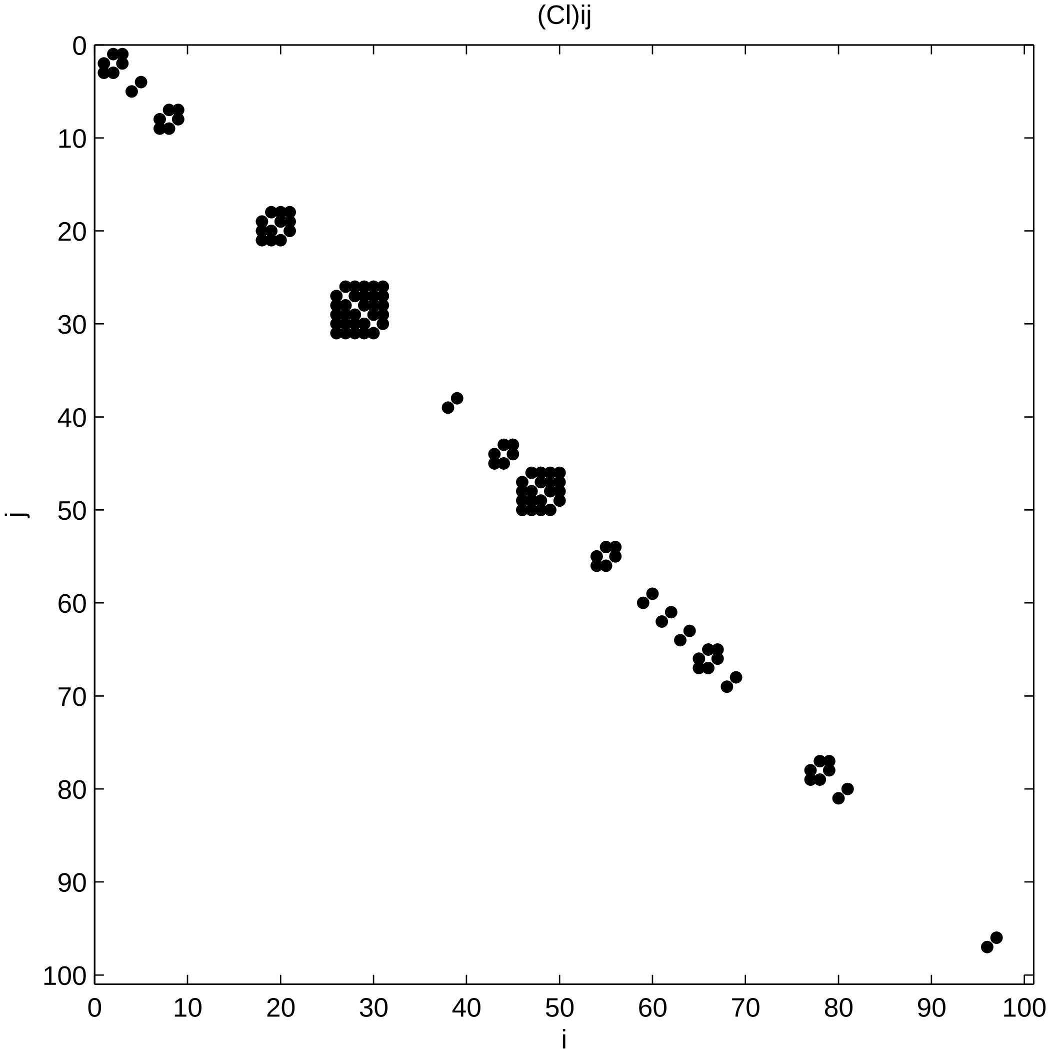


FIG. 3: Cluster connectivity matrix for reshuffled connectivity matrix C.

distance between them.

Some geometrical aspects of the $n-1$ simplex and its $p-1$ dimensional sub-simplex faces are relevant to our dynamical evolution. Most such features can be derived without utilizing any specific coordinate representation.

The fundamental relation

$$\vec{r}_i \cdot \vec{r}_j = -\frac{1}{n-1} \qquad all \quad i \neq j \quad i,j = 1 \ldots n \tag{A1}$$

was derived above by using $(\sum \vec{r}_i)^2 = 0$ and symmetry. It allowed us to deduce the length of any edge

$$|\vec{r}_i - \vec{r}_j| = \sqrt{\frac{2n}{n-1}} \qquad all \quad i \neq j \tag{A2}$$

such edges can be viewed as 1-dim 2 point subsimplices.

We have also triangles, namely 2-simplices, forming 2-dim "faces"/edges etc, $\begin{pmatrix} n \\ p \end{pmatrix}$ different $p-1$-simplices etc.

Let r_p denote the radius of the sphere circumscribing the $p-1$ simplex and d_p the distance to its center from the origin (namely the center of the original $n-1$ simplex). Clearly $d_p^2 + r_p^2 = 1$.

Let $\vec{r}_{i_1} \ldots \vec{r}_{i_p}$ be the p unit vectors of the p simplex. All the i_p are different and there are $\begin{pmatrix} n \\ p \end{pmatrix}$ such possible subsets of the n original $\vec{r}_i$. The vector from the origin to the center of simplex is:

$$\vec{d}_p = (\vec{r}_{i_1} + \vec{r}_{i_2} + \ldots \vec{r}_{i_p})/p \tag{A3}$$

Hence using again (A1) we find:

$$d_p = \sqrt{\vec{d}_p^2} = \frac{1}{p}\sqrt{p - \frac{p(p-1)}{n-1}} = \sqrt{\frac{n-p}{(n-1)p}}. \tag{A4}$$

And

$$r_p = \sqrt{1-d_p^2} = \sqrt{\frac{n}{n-1} \cdot \frac{p-1}{p}}, \tag{A5}$$

r_p is the distance from vertex of the p simplex to its center. Except for very small p's (representing "tiny" clusters) all r_p are $O(1)$ so formation of such clusters would require the vertices to travel the same distance as in the formation of bigger clusters.

The actual angular separation between $\vec{r_i}$ in the p simplex and r_p, the vector from the origin to its center is given by:

$$\theta_p = \arccos{(d_p)} = \frac{\pi}{2} - \arcsin{(r_p)} \approx \frac{\pi}{2} - \sqrt{\frac{n-p}{(n-1)p}}. \tag{A6}$$

Two p simplices can differ by just one, two $\ldots q, \ldots$ or $p-1$ points. The distances r_p^q between the centers of two neighboring p simplices differing by q vertices (and with $p-q$ common vertices) grow with q for fixed p, as follows.

The vector connecting the two centers is:

$$\vec{d_p} = \frac{1}{p}\left(\sum_{i=1}^{q} \vec{r_i} - \sum_{i=1}^{q} \vec{r_{li}}\right). \tag{A7}$$

With the sets $\{\vec{r_i}\}$,$(\{\vec{r_{li}}\})$ denoting the q points in the first (second) p simplices which are not shared by the two. The common $\vec{r_i}$'s cancel in the difference, and do not contribute to the distance r_p^q. Using (A7) and $\vec{r_i} \cdot \vec{r_j} = -1/(n-1)$ we find:

$$r_p^2 = \sqrt{(\vec{d_p^q})^2} = \frac{1}{p}\sqrt{2q + \frac{2q}{n-1}}. \tag{A8}$$

Hence the angle between $\vec{r_p}^{(1)}$ the $\vec{r_p}^{(2)}$ vectors to the centers of the two simplices is given by:

$$\theta_p^q = 2\arcsin\left(\frac{r_p^q}{2d_p}\right) = 2\arcsin\left[\sqrt{\left(\frac{n}{n-p}\right)\frac{q}{2p}}\right]. \tag{A9}$$

The last equation displays a nice feature. There is a small angular distances between the centers of the (spherical) faces corresponding to two p clusters which differ by a small fraction q/p of their vertices. The angular separation grows once $q \approx p \ll n$ to $\theta_p^q \approx \pi/2$.

For our simulations we need an explicit representation of $\vec{r_i}$. Assume we know the latter for the $n-1$ simplex (in $n-2$ dimension) denote them by $\vec{\vartheta}_1 \ldots \vec{\vartheta}_{n-1}$ with each $\vec{\vartheta}$ an $n-2$ vector with known components:

$$\vec{\vartheta_j} = \vec{\vartheta}_{j1}\hat{e}_1 + \ldots \vec{\vartheta}_{j,n-2}\hat{e}_{n-2} \tag{A10}$$

with $\hat{e}_l$ the unit vector along the l-th axis. When $n-1 \rightarrow n$ we choose

$$\begin{aligned} \vec{r_n} &= \hat{e}_{n-1} \\ \vec{r_i} &= \lambda_n \vec{\vartheta_i} - \frac{1}{n-1}\hat{e}_{n-1} \quad i = 1 \ldots n-1. \end{aligned} \tag{A11}$$

The normalizing factor $\lambda_n = \sqrt{1 - 1/(n-1)^2}$ ensures $|\vec{r_i}| = 1$, given that $|\vec{\vartheta_i}| = 1$. Thus, starting with an one simplex with $x_1^1 = 1$; $x_2^1 = -1$, we inductively generate any $n-1$ simplex.

ACKNOWLEDGMENTS

S.Nussinov would like to dedicate this work to sir Isac Wolfson who donated the chair in theoretical physics at Tel-Aviv University on the occasion of his 80th birthday.

[1] T. Sudkamp, "Languages and Machines: An Introduction to the Theory of Computer Science", Addison-Wesley, 1997.

[2] D. Cvetković, P. Rowlinson and S. Simić, "Eigenspaces of graphs" (Encyclopedia of mathematics and its applications, v. 66), Cambridge; New York : Cambridge University Press, 258p., 1997.

[3] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **84**, 7524 (1987).

[4] J. D. Bryngelson and P. G. Wolynes, J. Phys. C **93**, 6902 (1989).

[5] T. Kennedy, J. Statist. Phys. **106**, 407 (2002).

[6] A. Jaeckel and J. Dayantis, Macromol. Theory Simul. **10**, 461 (2001).

[7] An analogous physical system was used by Farhi, Goldstone and Gutmann and Sipser arXiv quant-ph/0001106 (2000). Their idea was to create a eave function of n spins satisfying a set of Boolean logic logic requirments via adiabatic changing of the Hamiltonian.

[9] Due to Cayley-Hamilton theorem C^n can be expressed via the characteristic polynomial in terms of lower powers of C.

[] The task of finding completely self-avoiding paths (walks) and eventually Hamiltonian paths which never revisit anyvertex twice and trace through all the vertices ones, is much more demanding and cannot be archived via these single techniques. Indeed exact enumeration of all SAW (self-avoiding walks) on even very regular graphs corresponding to regular latices yields exact solutions of the Ising models in the corresponding dimensionalities (but not vice versa[5, 6]). The Hamiltonian part (cycle) problem is known to be a truly hard NP complete problem [1].

[] It is amusing to note - though this has no impact on the present work - that such forces are believes to arise in QCD and cause confinement there.

[] This inverse hierarchic is indeed happen in cosmological structure formation.

Appendix B

Graph equivalence and characterization via a continuous evolution of a physical analog

Vladimir Gudkov[1,*] and Shmuel Nussinov[2,†]

[1]Department of Physics and Astronomy
University of South Carolina
Columbia, SC 29208
[2]Tel-Aviv University
School of Physics and Astronomy
Tel-Aviv, Israel
and
Department of Physics and Astronomy
University of South Carolina Columbia, SC 29208

(Dated: January 27, 2005)

Abstract

A general novel approach mapping discrete, combinatorial, graph-theoretic problems onto "physical" models - namely n simplexes in $n-1$ dimensions - is applied to the graph equivalence problem. It is shown to solve this long standing problem in polynomial, short, time.

INTRODUCTION

A graph G consists of n vertices V_i connected by edges E_{ij}. It is described by a connectivity matrix C with:

$$C_{ij} = C_{ji} = 0, 1 \quad (for\ (dis)connected\ V_i\ and\ V_j\ i \neq j = 1, \cdots, n)$$

$$C_{ii} = 0 \tag{1}$$

Vertex relabelling $i \rightarrow p(i)$ leaves G invariant but changes C according to

$$C \longrightarrow C' = P^T C P \tag{2}$$

with P an orthogonal matrix with only one non-zero element in each row i and column $j = p(i)$, which represents the above permutation

$$P = \delta_{(j,p(i))} \tag{3}$$

The graph equivalence problem is the following: "Given C and C', how can we decide, in time which is polynomial in n, if both correspond to the same topological graph G or to different graphs?, or stated differently, does a permutation matrix P for which Eq.(2) holds exist, and what is this P matrix?"

Exhaustive testing of all $n!$ permutation is impractical even for moderate n. A more systematic search of P performs just those transpositions which enhance an "overlap" function say

$$trC^T C' = \sum_{ij} C_{ij} C'_{ij} \tag{4}$$

However the changes in C (and in $trC^T C'$) due to any permutation is finite. There is no algorithm for systematically enhancing $trC^T C'$, as subsequent transpositions may undo the improvement due to previous permutations.

Our basic suggestion is: Instead of using discrete, large changes of say just two elements in a transposition $(i \leftrightarrow j)$, we modify, in each step, all elements by small amounts.

Such "continuous" changes seem impossible: in the strict formal approach there are no "continuous permutations".

THE DYNAMICAL MODEL FOR SIMPLEX DISTORTION

We use a symmetric n simplex (in $n-1$ dimensions) to represent our graph. The "abstract" vertices V_i of G (or C) are mapped into the geometrical vertices $\vec{r}_i$, $\quad i = 1, \cdots, n$ of the simplex. The symmetric configuration with all $|\vec{r}_i - \vec{r}_j| \quad i \neq j = 1, \cdots, n$, equal, is the starting point of our algorithms.

The motion generated by the dynamics, was designed to distort the simplex by shifting its vertices from the symmetric initial positions. The distorted simplex then reveals characteristic features of the graph G [1].

The original aim of the distortion algorithm was to find groups of vertices in G with higher than average mutual connectivity, and asses the distances between the various clusters in the graph.

To this end attractive (repulsive) interactions were introduced between fictious point objects at $\vec{r}_i$ and $\vec{r}_j$ when the corresponding vertices V_i and V_j are connected (or disconnected) in G. We use first order "Aristotelian" dynamics:

$$\mu \frac{d\vec{r}_i(t)}{dt} = \vec{F}_i(\vec{r}_i(t)), \tag{5}$$

with forces $\vec{F}_i$ which derive from potentials:

$$\vec{F}_i = -\vec{\nabla}_{(\vec{r}_i)} \{ U \left[\vec{r}_i, \cdots, \vec{r}_n \right] \}, \tag{6}$$

$$U = \sum_{i>j} U_a(|\vec{r}_i - \vec{r}_j|) C_{ij} + \sum_{i>j} U_r(|\vec{r}_i - \vec{r}_j|)(1 - C_{ij}), \tag{7}$$

$U_a(r)$ ($U_r(r)$) are attractive (repulsive) pair-wise potentials.

By a proper tuning of the latter- which can even be modified as a function of "time" - we can physically cluster at separate locations groups of points representing strongly (internally) connected clusters in the graph G.

To avoid collapse towards the origin (or a "run-away" to infinity) if U_a (or U_r) dominates, we force $\vec{r}_i(t)$ to stay, at all times, on the unit sphere:

$$|\vec{r}_i(t)| = 1 \qquad all \quad t > 0. \tag{8}$$

The graph characterization (G.C.) and graph equivalence (G.E.P.) problems are very closely connected. If we could find (in polynomial number of steps!) a set of real numbers $\rho_1, \rho_2, \cdots, \rho_m$ that would completely characterize a graph G then the G.E.P is readily solved.

All we need to do is to compute for C (C') these numbers $\{\rho_k\}$ ($\{\rho'_k\}$), order the ρ_k and ρ'_k sets separately and compare them.

The set of eigenvalues $(\lambda_1, \cdots, \lambda_n)$ of the connectivity matrix are certainly invariant under relabelling. While this set encodes a rich body of information of graph theoretic interest, it fails to completely characterize graphs[2].

An alternative and natural simple variable helping characterize connectivity matrices is the mutual entropy (see, for example [3]). Suppose the connectivity matrix C has been normalized so that

$$\sum_{i,j=1}^{n} C_{ij} = 1. \tag{9}$$

$P_i = \sum_j^n C_{ij}$ could then be considered as the probability that V_i and V_j are connected. The corresponding entropy

$$H(row) = -\sum_{j=1}^{n} P_i \log P_i, \tag{10}$$

could be considered as a measure of the uncertainty of the rows connection for the given network. The amount of uncertainty for the connection of the column nodes given that the row nodes are connected is

$$H(column|row) = -\sum_{i,j}^{n} C_{ij} \log C_{ij} - H(row). \tag{11}$$

As a result the amount of *mutual information* gained via the given connectivity of the network is

$$I(C) = H(row) + H(column) - H(column, row), = \sum_{i,j}^{n} C_{ij} \log\left(C_{ij}/P_i P_j\right). \tag{12}$$

where

$$H(column, row) = -\sum_{i,j}^{n} C_{ij} \log\left(C_{ij}\right). \tag{13}$$

Due to the double summation and the symmetry of the connectivity matrix $I(C)$ does not depend on the vertex relabelling and is a permutation invariant measure for the connectivity matrix.

Calculations of the mutual entropy for two connectivity matrices provides an easy way to distinguish between these corresponding different graphs. If, however, the entropies are the same, the more detailed approach below is used. Amusingly we found that the entropy is already sufficient to distinguish between the lowest cospectral graphs (see, for example [4] and references therein).

The distances between the various vertices

$$r_{ij}(t) \equiv |\vec{r}_i(t) - \vec{r}_j(t)| \tag{14}$$

vary in our original algorithm as a function of time away from the original common value:

$$r_{ij}(0) = |\vec{r}_i(0) - \vec{r}_j(0)| = a \qquad all\ i \neq j = 1, \cdots, n \tag{15}$$

Also in identical simulations of the dynamical evolution, the sets of relative distances computed for C and C', should be the same if C and C' are equivalent:

$$\{r_{ij}(t)\} = \{r'_{ij}(t)\} \tag{16}$$

One permutation of n elements (namely that which brings via Eqs.(2) and (3) C into C') should yield:

$$|\vec{r}_{p(i)}(t) - \vec{r}_{p(j)}(t)| = |\vec{r'}_i(t) - \vec{r'}_j(t)| \tag{17}$$

It is straightforward to verify (16) and then using (17) recover the permutation $i \rightarrow p(i)$.

In essence the idea of the present algorithm is to use the distortion of the simplex $S(0) \rightarrow S(t)$ {i.e. $\vec{r}_i(0) \rightarrow \vec{r}_i(t)$} generated via the dynamics of (repulsion) attraction between (dis)connected vertices in G to bring out an "intrinsic shape" of the graph.

Initially all vertices were at equal distances[5]. All the information pertaining to the graph was encoded in the interactions of Eq.(7).

After enough evolution steps, each vertex moves appreciably away, namely by

$$|\vec{r}_i(t) - \vec{r}_i(0)| \approx a/2 \tag{18}$$

from its initial position. The information on the specific graph G reflects in the geometrical shape of S, i.e. the set of distances,

$$|\vec{r}_i(t) - \vec{r}_j(t)| \qquad i \neq j = 1, \cdots, n. \tag{19}$$

Vertices which are near in a graph theoretic sense, namely for which there are many, short, connecting paths in the graph move closer together. (A short path consists of a small # of consecutive edges which starts at V_i say and terminates at V_j). Like wise vertices which are far in a graph theoretic sense i.e. have fewer and longer connecting paths will tend to move further away.

In our earlier work[6] we sought to identify "clusters in the graphs" namely have the points corresponding to a subset $\{C_i\}$ of vertices in the graph which have relatively strong mutual, internal, connectivity, collapse to a single point.

For the present purpose we need (and should!) not pursue the evolution that far, as by then the graph simplifies and some of the inter-cluster details are lost. Rather we need to stop "Half-Way": after Eq.(19) holds and yet no cluster has completely collapsed.

Note that in $n-1$ dimensions all the $n(n-1)/2$ distances $|\vec{r}_i(t) - \vec{r}_j(t)|$ are independent, apart from triangular inequalities of the form

$$|\vec{r}_i(t) - \vec{r}_j(t)| \leq |\vec{r}_i(t) - \vec{r}_k(t)| + |\vec{r}_k(t) - \vec{r}_i(t)|. \tag{20}$$

Jointly these distances specify the geometric shape of S.

The mapping of the $n(n-1)/2$ bits of information: $C_{ij} = 0$ *or* 1, via our dynamic evolution, into the set of $n(n-1)/2$ distances, is highly non-linear. The fact that we have $n(n-1)/2$ distances (rather than just n eigenvalues) makes the former more likely to specify the graphs.

Further we note that the time t when the comparisons are made and the attractive and repulsive interactions in Eq.(7) above are free parameters and functions. Hence we can repeat the above graph comparisons for many values and/or many functions $U_r(\rho)$, $U_a(\rho)$, making the significance of a successful match extremely high.

If many of the $r_{ij}(t)$ {and $r'_{ij}(t)$} are degenerate our ability to resolve graphs will be diminished. However such degeneracies must stem from some symmetries in the graphs and corresponding connectivity matrices. Once these symmetries are identified, the number of independent C_{ij} (or C'_{ij}) and the task of comparing them will be accordingly reduced.

We apply the above approach below, demonstrating its power and versatility.

THE CONVERGENCE AND COMPLEXITY OF THE DISCRETE MODELINGS OF THE DYNAMICAL EVOLUTIONS

We follow the dynamics of the vertex shifts in Eq.(5) by discretizing the first order equations:

$$\vec{r}_i(t+\delta) = \vec{r}_i(t) + \frac{\delta}{\mu}\vec{F}_i(\vec{r}_e(t)) \tag{21}$$

with δ a small time increment.

Since $\vec{r}_i$, $\vec{F}_i$, are $n-1$ dimensional vectors Eq.(21) represents $\mathrm{O}(n^2)$ equations for the relevant components. Each force component $F_{i\alpha}$ is a sum of v_i force components with v_i the valency of the vertex V_i i.e. the # of vertices connected to it. Hence each step in (26) involves $n^2v/2$ calculations with

$$v = \sum_{i=1} v_i/n \tag{22}$$

the average valency in the graph.

Let us assume that we need to repeat the process of iterating the dynamics namely (26) or (27) for s steps in order to achieve the goal(s) of the algorithm(s). These goals vary for the various problems of interest. For cluster identification we need the points representing clusters in the graph to physically converge into definable separate regions.

For graph characterizations and comparison we need a fewer number of steps, sufficient to make the distances $r_{ij}(t)$ vary considerably away from their original common value.

The total number of computations involved is $N = O(n^2s)$ if v is finite and n independent or $N = O(n^3s)$ for the extreme case when $v \approx n$. For N not to be polynomial in n we need that s will grow faster than any power of n.

In principal one can envision many types of chaotic dynamical evolution where such large number of steps is indeed required.

This is not the case for the first order equations considered here:

$$\dot{\vec{r}_i} = \frac{\vec{F}_i}{\mu} = -\frac{\vec{\nabla}_{r_i}(U)}{\mu}, \tag{23}$$

where the system consistently moves, along the steepest descent, to a minimum of U, the potential energy.

If we have a complicated "energy landscape" the system can be trapped in any one of the many local minima, a feature which accounts for the difficulty of protein folding[7], neural nets and spin glass problems[8]. The need to keep the same deterministic evolution for S and S' representing G and G' in the first "distortion" algorithm, excludes in our case- the possibility of introducing some stochastic noise to extricate the system from a local minimum.

Fortunately our problem does not allow for many minima. Thus let us fix the locations of all $\vec{r_i}$ $\quad i = 1, \cdots, n-1$ except $\vec{r_n} \equiv \vec{r}$. The velocity $\dot{\vec{r}}(t)$, is dictated by

$$U(\vec{r}) = \sum_{i=1} C_{n_i} U_A(|\vec{r} - \vec{r_i}|) + (1 - C_{n_i}) U_R(|\vec{r} - \vec{r_i}|) \tag{24}$$

Assume we have some local equilibrium at $\vec{r_0}$. Locally, in the neighborhood of $\vec{r_0}$, we can use the variables $\rho_i \equiv |\vec{r} - \vec{r_i}| \quad i = 1, \cdots, n-1$, instead of $x_1 \cdots x_{n-1}$ the $n-1$ Cartesian coordinates of $\vec{r}$. The conditions for an extremum $\vec{\nabla} U(\vec{r}) \mid_{\vec{r}=\vec{r}_o}$ then require that

$$\frac{\partial}{\partial \rho_j} U_A(\rho_j) \mid_{\rho_j = \rho_j^{(o)}} = 0; \quad or \quad \frac{\partial}{\partial \rho_j} U_R(\rho_j) \mid_{\rho_j = \rho_j^{(o)}} = 0 \tag{25}$$

Thus for generic monotonic U_A, U_R, we have no extrema inside the region.
An absolute minimum obtained at the boundary.

APPLICATIONS OF THE METHOD

To demonstrate the power of our approach we considered a graph with 100 vertices each of which is randomly connected to seven others. The corresponding connectivity matrix C is shown in Fig.(1)). Random reshuffling transforms the C into the matrix B of Fig.(2) Next we applied our algorithm using a combination of attractive and repulsive forces in $n - 1 = 99$ dimensional space. The vertices of the 100-simplex were allowed to move under the influence of the forces on the 98-dimensional hyper-sphere in 99-dimensions. After a number of steps we analyzed the distances between pairs the vertices of the two simplexes. We found perfect correspondence between the distance matrices. We also readily show the permutation matrix which maps one distance matrix on to the another. Applying the latter to the matrix B reproduces exactly the original connectivity matrix C (Fig.(1)).

S.Nussinov would like to thank Zohar nussinov for a crucial comment regarding the advantage of going to higher dimensions to overcome frustrations and alleviate constraints.

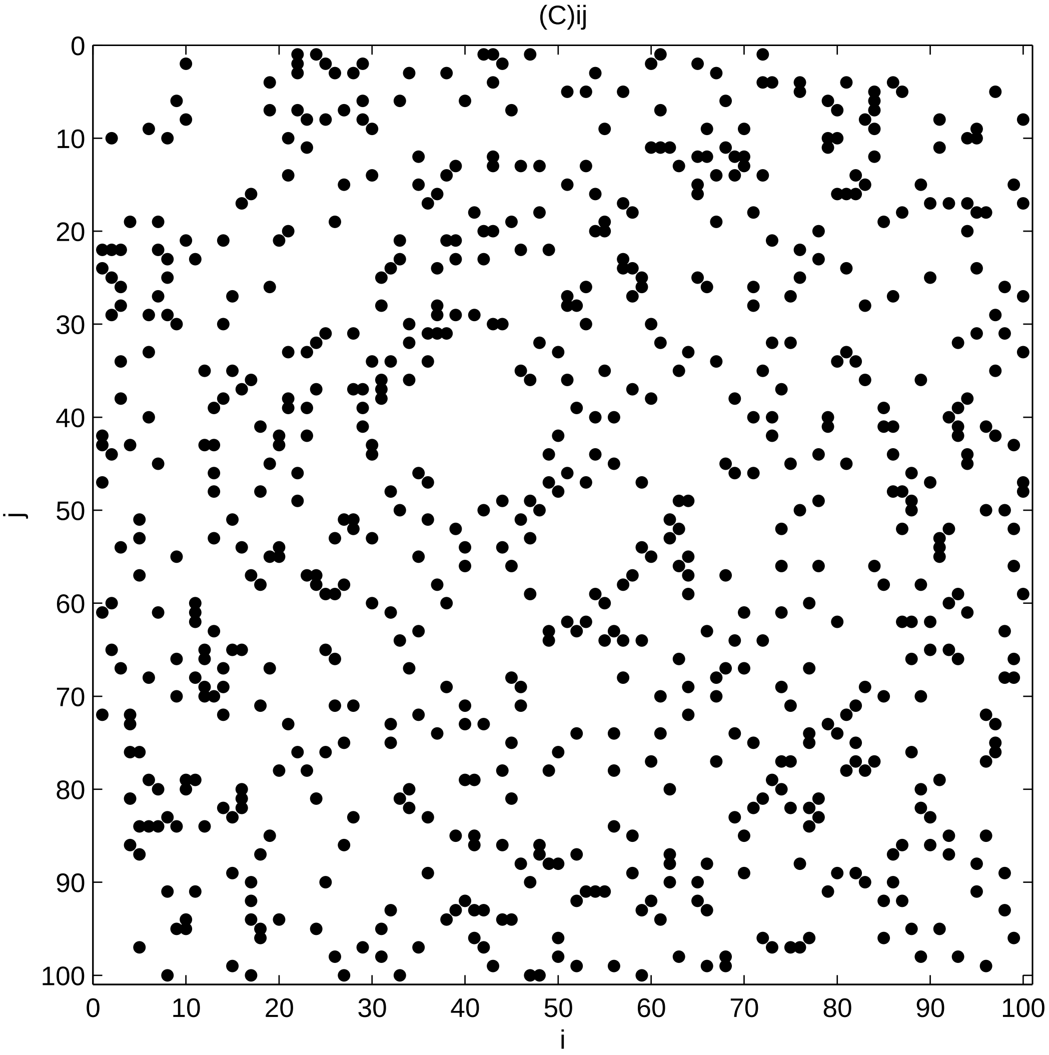


FIG. 1: Connectivity matrix for 100 vertices graph with 7 random connections for each vertex.

He would like to dedicate this work to sir Isac Wolfson who donated the chair in theoretical physics at Tel-Aviv University on the occasion of his 80th birthday.

[*] `gudkov@sc.edu`

[†] `nussinov@ccsg.tau.ac.il`

[1] An analogous physical system was used by Farhi, Goldstone and Gutmann and Sipser arXiv quant-ph/0001106 (2000). Their idea was to create a eave function of n spins satisfying a set of Boolean logic logic requirments via adiabatic changing of the Hamiltonian.

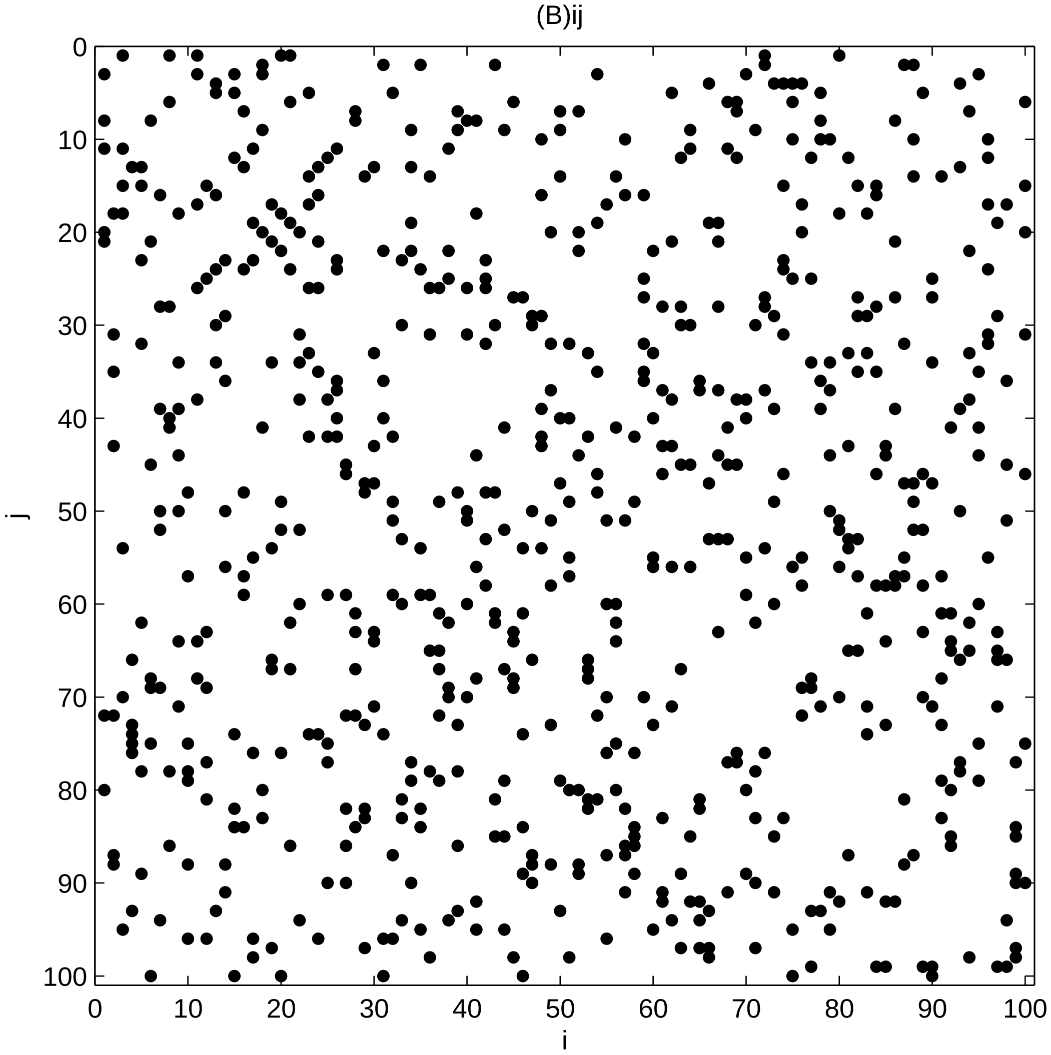


FIG. 2: Reshuffled connectivity matrix.

[2] D. Cvetković, P. Rowlinson and S. Simić, “Eigenspaces of graphs” (Encyclopedia of mathematics and its applications, v. 66), Cambridge; New York : Cambridge University Press, 258p., 1997.

[3] A. Rényi, “Probability Theory”, North-Holland Publishing Company - Amsterdam - London, and American Elsevier Publishing Company, Inc. - New York, 1970.

[4] C.D.Godsil, D.A. Holton and B.D. McKay, “The spectrum of a graph”, Lecture Notes in Math. **622**, Springer-Verlag, Berlin, 1977,91-117.

[5] For a specific convenient coordinate choice for the n vertices see Appendix of [6].

[6] V. Gudkov, J.E. Johnson and S. Nussinov, arXiv: cond-mat/0209111 (2002).

[7] J.D. Bryngelson and P.G. Wolynes, J. Phys. Chem. **93**, p. 6902 (1989).

[8] M. M'ezard, G. Parisi and M.A. Virasoro, "Spin glass theory and beyond", World Scientific, Singapore, 1987.

Appendix C

A Novel Approach Applied to the Largest Clique Problem

Vladimir Gudkov[*]

Department of Physics and Astronomy
University of South Carolina
Columbia, SC 29208

Shmuel Nussinov[†]

Department of Physics
Johns Hopkins University
Baltimore MD 21218
and
Tel-Aviv University,
School of Physics and Astronomy
Tel-Aviv, Israel

Zohar Nussinov[‡]

Institute Lorentz for Theoretical Physics,
Leiden University
POB 9506, 2300 RA Leiden,
The Netherlands

(Dated: January 27, 2005)

Abstract

A novel approach to complex problems has been previously applied to graph classification and the graph equivalence problem. Here we apply it to the NP complete problem of finding the largest perfect clique within a graph G.

PACS numbers: 89.75.Hc, 89.90.+n, 46.70.-p, 95.75.Pq

INTRODUCTION

In a novel dynamical approach the n vertices of a graph G are mapped onto n physical points located initially at equal distances from each other forming a symmetric n simplex in $n-1$ dimensions. Attractive/repulsive forces are introduced between pairs of points corresponding to connected/disconnected vertices in the original graph G. We then let the system evolve utilizing first order Aristotelian dynamics[1–3]. We tune the relative strength of repulsive and attractive forces to be v/n with v the average valency i.e. average number of vertices connected to a given vertex so as to have no net average repulsion/attractions.

We found, that as the system evolves various physical clusters of points tend to form. These physical clusterings reveal clusters (or imperfect cliques) in the graph - namely groups of vertices with a larger than average mutual connectivity. Also the matrix of distances $R_{ij}(t)$ between the various points $\vec{r}_i(t)$ and $\vec{r}_j(t)$ is characteristic of the graph topology: points corresponding to vertices which are "close in the graph" namely have (relatively) many, short, paths connecting them will move closer together and conversely, points which are "far in the graph" tend to move apart.

The distance matrix and clusters are important graph diagnostics. In particular the first allows us to solve easily the graph equivalence problem namely to decide if two connectivity matrices C_{ij} and $C\prime_{ij}$ correspond to the same topological graph and if they do to find the relabelling of vertices which makes C and $C\prime$ identical.

These results are of considerable practical importance. Still neither of the above problems belongs in the special class "NP complete" problems. The latter consists of problems such as the travelling salesmen problem and the satisfiability problem for which a putative solution can be readily checked in polynomial time yet no polynomial solution method is presently known[4].

Many of these problems can be phrased in terms of graphs as the task of finding some specific graph g inside bigger graph G. Further, all these problems which superficially seem very different are at a basic level ,equally difficult: If a method of solution in polynomial time is found for one such problem then all the problems should be solvable in such time by essentially the same method. Conversely if we can prove that just one NPC problem necessarily require, non-polynomial time for its solution, the same holds for all of them. Two of us have recently conjectured[3] that a new variant of our approach namely of dynamically docking rigid simplexes s and S representing g and G can solve the "g inside G" problems. Here we wish to present the first concrete application of the original, point translation or single simplex distortion algorithm (SDA), to an NPC problems namely that of finding the largest perfect clique in G.

To most clearly illustrate the essence of the problem we consider the "students in dorm" example used in the general description of the Clay institute prize offered for resolving $P = NP$ problem[5]. We have $N = 400$ students out of which we need to select $n = 100$ which can live together in a dorm, subject to a very long list of mutual exclusions. This list states that student # 1 cannot be together with any one student from a specific set of say 200 other students, student student # 2 cannot be together with any one from another partially overlapping set with a comparable number of students etc. How can we pick up a set of 100 students such that any one is completely compatible with the other 99, and what is this set? Clearly this is a particular example of the general satisfiability problem where the conditions imposed are just "two body" exclusions.

It is also a particular case of looking for a graph g inside G where g is a perfect clique of vertices each of which is connected to all other members in the clique. We encode into G with $N = 400$ vertices the various mutual exclusion constraints by not connecting with edges vertices V_i and V_j if student # i and student # j are not compatible, and connecting by edges compatible pairs. Clearly if we find within G a clique with n vertices it means, by our very construction, that the students to which these vertices correspond are indeed all mutually compatible. We could construct in judicious manner various smaller consistent subsets, and try piece them together. Often, however a new inconsistency is revealed and we need to pursue other alternatives. While we certainly can do this in far less steps than $\begin{pmatrix} N \\ n \end{pmatrix} = \begin{pmatrix} 400 \\ 100 \end{pmatrix}$ the difficulty of the problem seems to grow at least exponentially with n.

In desperation we might decide to resort to the following primitive alternative and simply let the 400 students "fight it out". In this all out war each student will try to push away members which are inconsistent with him and pull in those which are. This collective natural selection of the "compatible" - which may well be a prerelevant social phenomena - would hopefully leave us with the desired large group of mutually consistent individuals. Unfortunately the outcome of such a 400 way "Somo" fight of staying in the ring is strongly biased by the initial arbitrary placement of students in the two dimensional arena[6]. Thus we could envision a situation where an ideal group of completely compatible dorm candidates is placed in the center of a group of highly unpopular ones and is “ejected” together with them. In order to generate the correct large clique we need to completely unbias the starting position and avoid the severe constraints due to our existence in a physical world with limited number of dimensions. This can be done only if we go to $d = N - 1$ dimensions and place the “students” which, in the inverse problem that we are really after, are metaphors for the physical points representing the N vertices of the graph, at the vertices of a symmetric N simplex.

SEARCHING FOR CLIQUES

Our search for perfect cliques uses the same physically motivated dynamical algorithm previously developed to identify via the physically bunched points clusters or “imperfect cliques” in a graph[1]. We found that to adapt this algorithm for the present purpose we need only to enhance the ratio of the repulsive and attractive interactions. Originally it was chosen to be:

$$U_{rep}(r)/U_{att}(r) = v/n, \tag{1}$$

which could be relatively small. Thus for an average valency of 10 in a graph with 100 vertices it is only 0.1. However, in order to meet the criteria of *perfect* cliques we clearly have to significantly enhance the strength of the repulsive interactions so as to avoid points which are connected to a fairly large number of the points in the clique but not to ALL of them from joining in. Thus in the first round of applications we used

$$U_{rep}(r)/U_{att}(r) = 1. \tag{2}$$

To see how the algorithm works for the case of overlapping cliques we considered two cliques 7×7 and 15×15 with a 2×2 overlap on a “background” of a 100×100 matrix with the average 10% connectivity. The corresponding connectivity matrixes before reshuffling is shown in Fig.(1). To simulate a real-life situation of networks with unknown structure (topology) we randomly permute the rows and columns of the matrix C obtaining the reshuffled matrix C'. Next we apply our algorithm for clusters reconstruction using equal attractive and repulsive constant forces in $n - 1 = 99$ dimensional space. The vertices of the 100-simplex were allowed to move under the influence of the forces on the 98-dimensional hyper-sphere in 99-dimensions. After a number of steps we analyzed the mutual distances between the vertices of the simplex and group neighbors which are close to each other into cliques. The new cluster-connectivity matrix is shown in Fig.(2). The reconstructed

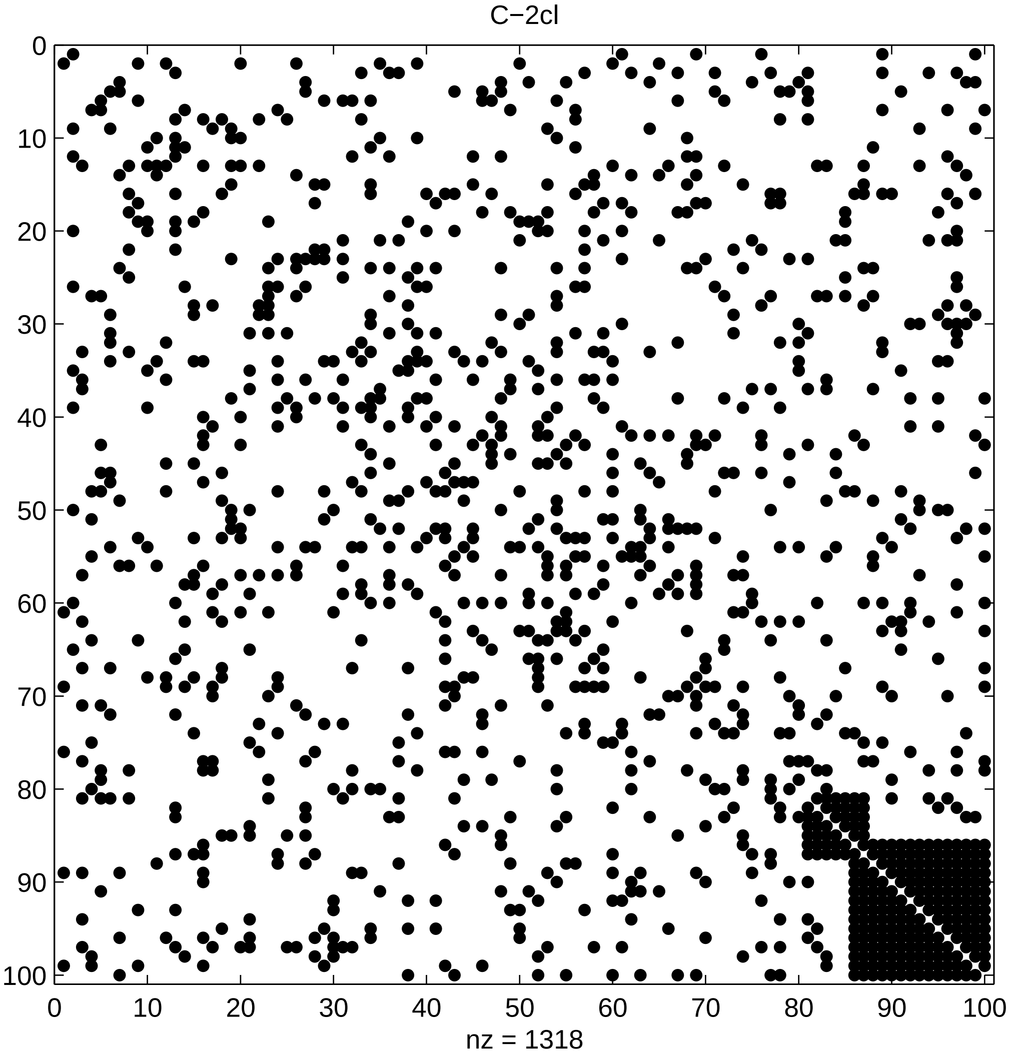


FIG. 1: Connectivity matrix C with 7×7 and 15×15 cliques.

connectivity matrix for the cliques is shown in Fig.(2). We see that due to the large repulsive forces most vertices did not move close each to others. The only vertices grouped together are the ones that belong to the cliques.

Other examples involve a $n = 100$ clique in a $N = 400$ graph corresponding to the "students in dorm" question. In addition we had an imperfect clique or cluster of 300 with average valency of 20% on a background of 10% (Fig.(3)). The 100 clique successfully reconstructed after reshuffling is shown in Fig.(4).

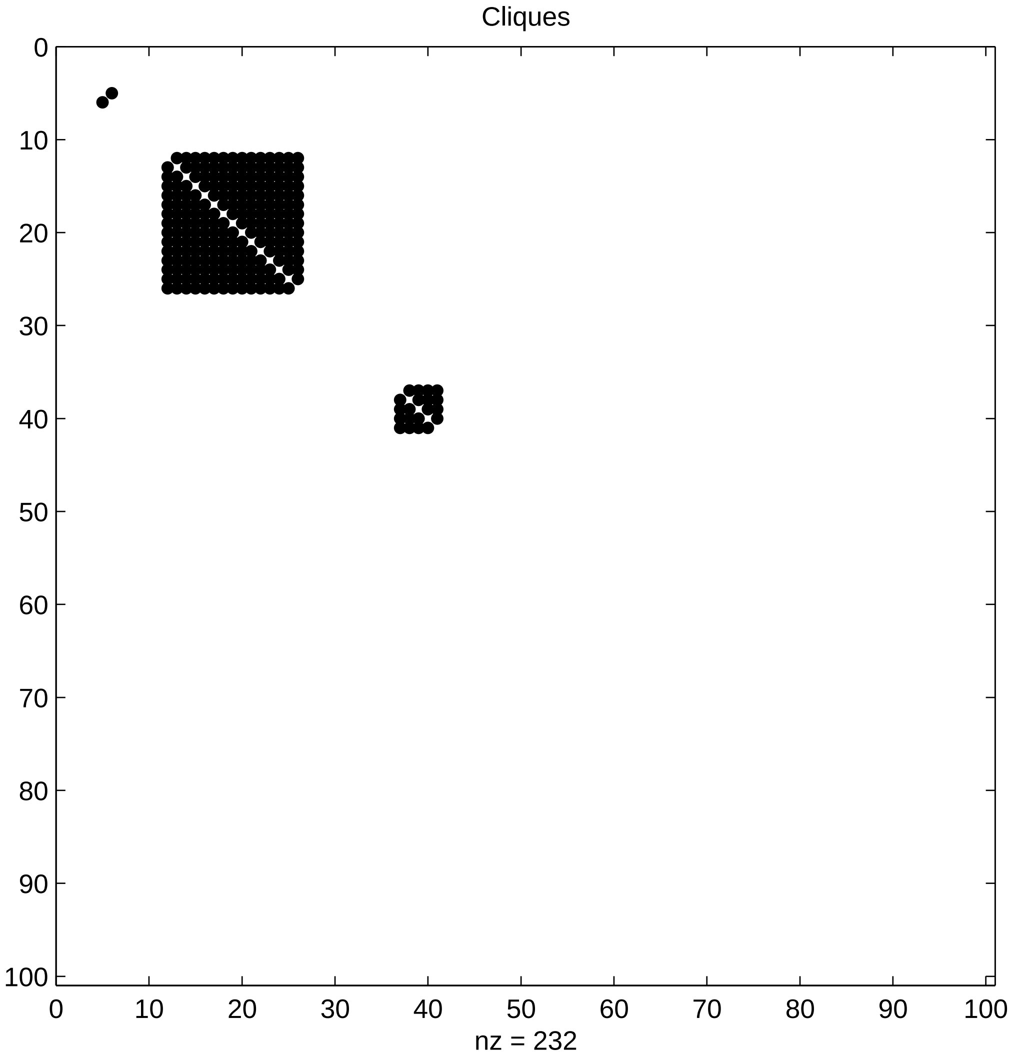


FIG. 2: Reconstructed clique connectivity matrix for $C\prime$ 7×7 and 15×15 cliques.

The above result is to our mind fairly impressive. It shows that our original code solves in very short time the NPC problem of the largest clique. It may still fall short of solving it in all cases. As a worst case scenario we could envision a vertex (or several such vertices) which are connected to all the vertices in the clique save one. To avoid these vertices from joining the clique even in this case, rendering it imperfect, we need that the single repulsion due to the missing edge, overcome all the $n-1$ attractions to the rest of the points in the clique . Thus the strict perfect clique worst case scenario demands

$$U_{rep} > (n-1) \cdot U_{att}. \tag{3}$$

This wildly differs from the above eq.(refx): for a graph G with 100 vertices $v = 10$ and a

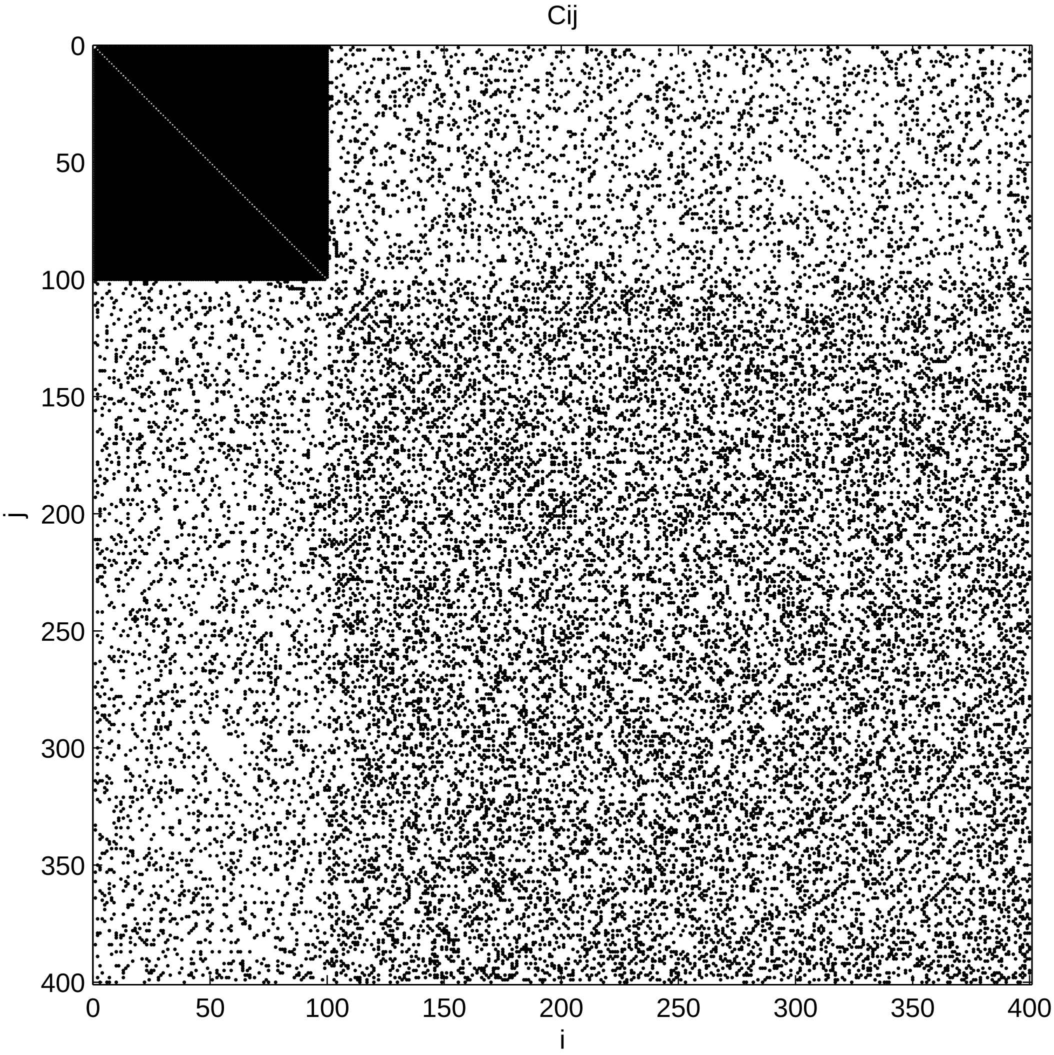


FIG. 3: Connectivity matrix C for 100×100 clique.

clique of size $n = 10$ we need a factor hundred enhancement of the ratio U_{rep}/U_{att} from 0.1 to 10!

Our $3 - d$ based intuition would strongly suggest that this stops formation of all cliques, perfect or not, since as any given point tries move towards its “Designated” clique it may be “Overwhelmed” by the many repulsive forces which will prevent it from joining the clique. The configuration with the perfect clique (and the largest perfect clique in particular) fully formed i.e having all its vertices collapse at a point is indeed the desired final lower energy state. However there may be false, local minima which trap our system just like in spin glass[7] and protein folding problem[8].

This is indeed most certainly the case for ”low” dimensionalities. However with $d = N-1$,

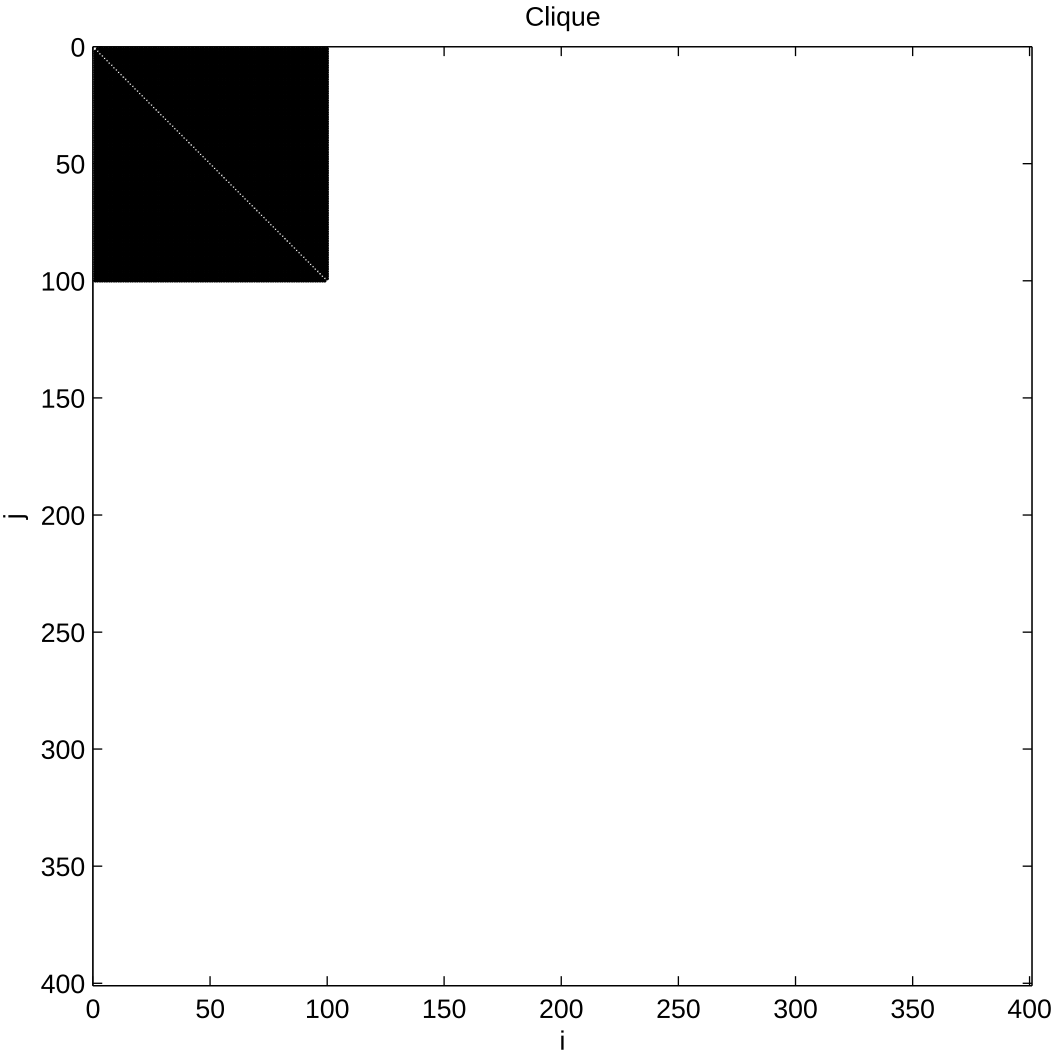


FIG. 4: Reconstructed clique connectivity matrix for 100 × 100 clique.

as is the case here, the above intuition fails. Specifically any one given "test point" feels just as many different forces in the directions of the other particles namely $N-1$ as there are independent directions $d = N-1$ to move in. Ideally therefore the test particle should be able to simultaneously respond to all different $N-1$ forces, move in the direction of all the attractors and away from all the repellers and in the process further lower the energy of the system. We can adopt a local, non-orthogonal, system of coordinates where the $N-1$ axes are aligned along the unit vectors pointing from $\vec{r}$ - the chosen point, to $\vec{r}_1, \vec{r}_2, \ldots \vec{r}_{N-1}$ the other $N-1$ points. Using our choice of constant forces[10] we have then a net force

$$F(\vec{r}) = \sum_{i=1}^{N-1} \frac{\vec{r} - \vec{r}_i}{|\vec{r} - \vec{r}_i|}, \tag{4}$$

which is the sum of the unit vectors along these axes with + and - signs. Since these are $N-1$ linearly independent vectors the sum never vanishes $|F(\vec{r})| > 0$ always and no local minimum arises.

There is one "small" correction however to the above argument. It is due to the fact that in our original algorithm we have introduced one further constraint on the motion of the points, namely that at all times on the unit circle $|\vec{r}_i(t)| = 1$. It seemed necessary in order to avoid running away to infinity of repelling vertices or collapse to the origin of attracting ones. This does however introduce an extra normal reaction force that could in fact cancel the above sum in eq.(4), and thus yields local minima. Hence in the final runs we did not impose this constraint. Instead we modified our code to facilitate handling the increasing distances between points at later stages of the evolution. We found that our program fully reconstructed the maximal clique[11]. This happens regardless of the degree of the connectivity of the random background and also of the existence of large and partially overlapping slightly smaller cliques. Thus for the n=100 maximal clique in an N=400 vertex graph (i.e the students choice for dorm problem) we added two 80x80 cliques which overlapped our 100x100 clique in two 60x60 patches which ,in turn, had a 20x20 overlap and used a background with 70% connectivity Fig.(5). Even under such seemingly unfavorable conditions we reconstructed our clique Fig.(6).

[*] gudkov@sc.edu

[†] nussinov@ccsg.tau.ac.il

[‡] Zohar@lorentz.leidenuniv.nl

[1] V. Gudkov, J.E. Johnson and S. Nussinov, arXiv: cond-mat/0209111 (2002).

[2] V. Gudkov and S. Nussinov, arXiv: cond-mat/0209112 (2002).

[3] S. Nussinov and Z. Nussinov, arXiv: cond-mat/0209155 (2002).

[4] T. Sudkamp, "Languages and Machines: An Introduction to the Theory of Computer Science", Addison-Wesley, 1997.

[5] Clay Institute, http://www.claymath.org/prizeproblems/pvsnp.htm.

[6] Social pressure / encouragement via nasty / favoureble communications are not limited by dimensionality-explaining its vast power and magical efficiency.

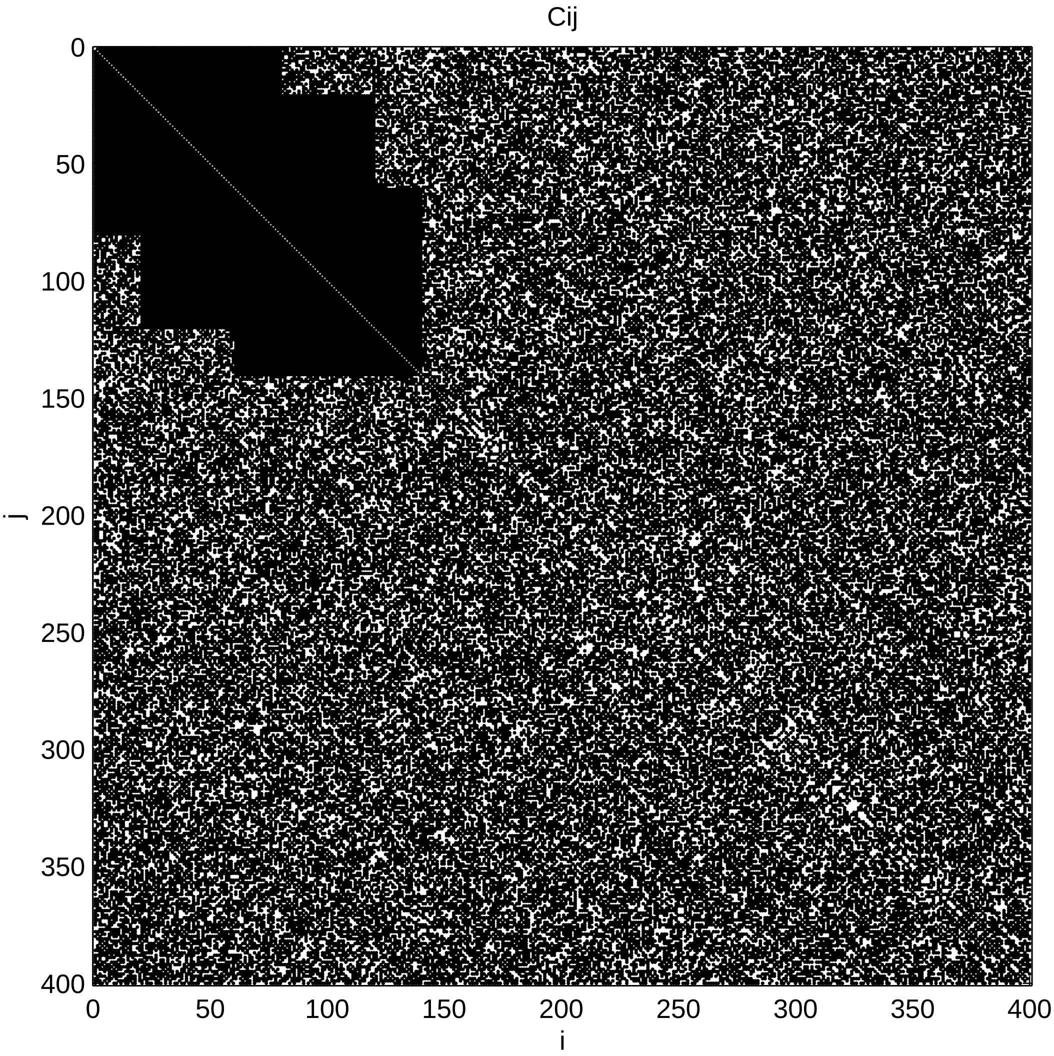


FIG. 5: Connectivity matrix with three overlapping cliques and 70% random background.

[7] M. M'ezard, G. Parisi and M.A. Virasoro, "Spin glass theory and beyond", World Scientific, Singapore, 1987.

[8] J.D. Bryngelson and P.G. Wolynes, J. Phys. Chem. **93**, p. 6902 (1989).

[9] An analogous physical system was used by Farhi, Goldstone and Gutmann and Sipser arXiv quant-ph/0001106 (2000). Their idea was to create a eave function of n spins satisfying a set of Boolean logic logic requirments via adiabatic changing of the Hamiltonian.

[10] This choice common to the present and earlier works, was originally made for simplicity. In retrospect it turns to be optimal. Indeed if the forces - or potentials become too strong at short distances,then two points which attract each other, may prematurely "seize" and become

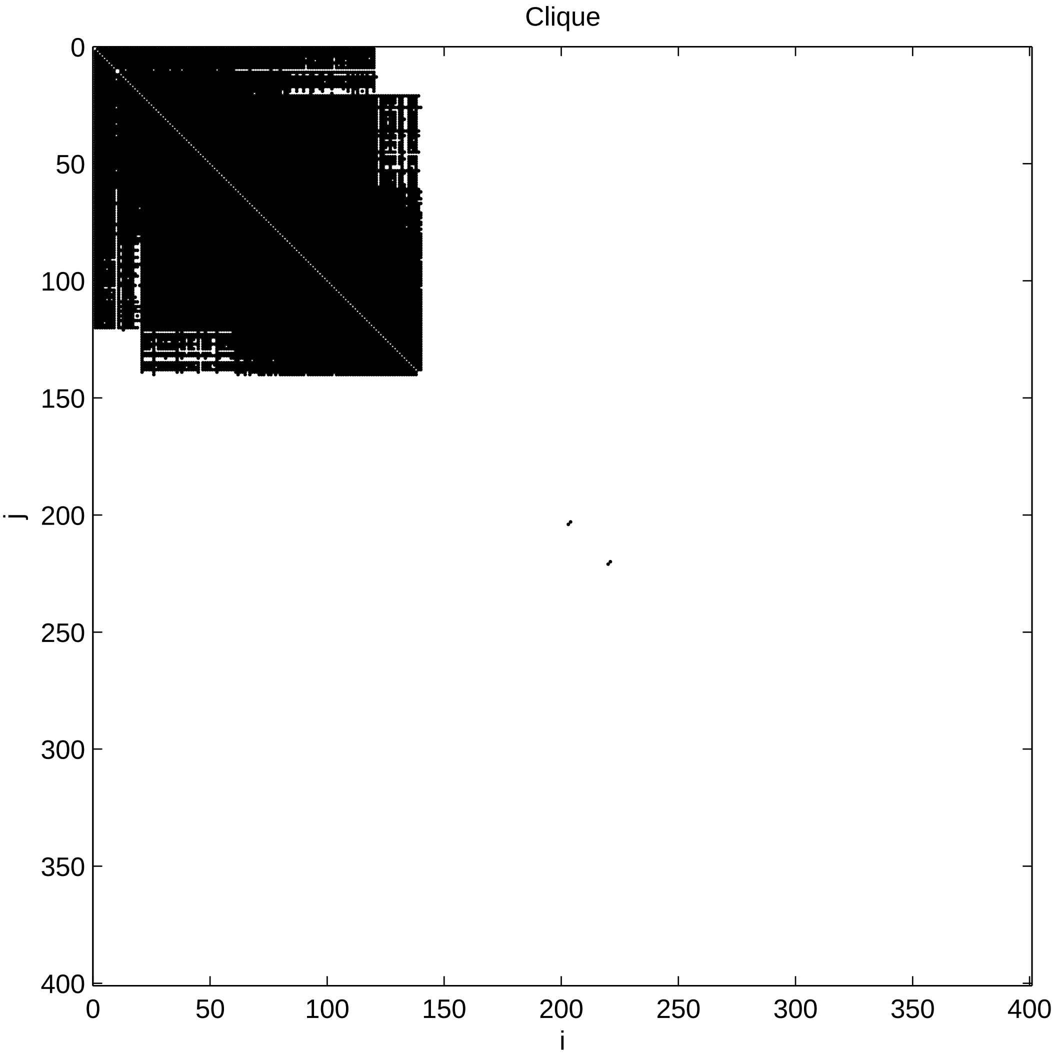


FIG. 6: Reconstructed cliques for 400-matrix with 70% background connectivity.

inseparable - a phenomenon which is a source of the deleterious multiple local minima. The general convexity condition for the potentials $U(r)$ which avoids this difficulty -and which our linear potential trivially satisfies- has been pointed to us by Vassilios S. Vassiliadis (private communication).

[11] Once we find a substantial portion $l < n$ of the vertices belonging in the perfect clique the construction of the rest of the clique can be easily completed as follows: We omit from our graph G all those vertices which do not connect to ALL the vertices in the clique. This will drastically reduce the number of vertices in the remaining relevant part of G and vastly accelerate the remaining calculations.

Appendix D

Multidimensional Network Monitoring for Intrusion Detection

Vladimir Gudkov and Joseph E. Johnson
Department of Physics and Astronomy
University of South Carolina
Columbia, SC 29208
gudkov@sc.edu; jjohnson@sc.edu

An approach for real-time network monitoring in terms of numerical time-dependant functions of protocol parameters is suggested. Applying complex systems theory for information flow analysis of networks, the information traffic is described as a trajectory in multi-dimensional parameter-time space with about 10-12 dimensions. The network traffic description is synthesized by applying methods of theoretical physics and complex systems theory, to provide a robust approach for network monitoring that detects known intrusions, and supports developing real systems for detection of unknown intrusions. The methods of data analysis and pattern recognition presented are the basis of a technology study for an automatic intrusion detection system that detects the attack in the reconnaissance stage.

1.1 Introduction

Understanding the behavior of an information network and describing its main features are very important for information exchange protection on computerized information systems. Existing approaches for the study of network attack tolerance usually include the study of the dependance of network stability on network complexity and topology (see, for example [1, 2] and references therein);

signature-based analysis technique; and statistical analysis and modelling of network traffic (see, for example [3, 4, 5, 6]). Recently, methods to study spatial traffic flows[7] and correlation functions of irregular sequences of numbers occurring in the operation of computer networks [8] have been proposed.

Herein we discuss properties related to information exchange on the network rather than network structure and topology. Using general properties of information flow on a network we suggest a new approach for network monitoring and intrusion detection, an approach based on complete network monitoring. For detailed analysis of information exchange on a network we apply methods used in physics to analyze complex systems. These methods are rather powerful for general analysis and provide a guideline by which to apply the result for practical purposes such as real time network monitoring, and possibly, solutions for real-time intrusion detection[9].

1.2 Description of Information Flow

A careful analysis of information exchange on networks leads to the appropriate method to describe information flow in terms of numerical functions. It gives us a mathematical description of the information exchange processes, the basis for network simulations and analysis.

To describe the information flow on a network, we work on the level of packet exchange between computers. The structure of the packets and their sizes vary and depend on the process. In general, each packet consists of a header and attached (encapsulated) data. Since the data part does not affect packet propagation through the network, we consider only information included in headers. We recall that the header consists of encapsulated protocols related to different layers of communications, from a link layer to an application layer. The information contained in the headers controls all network traffic. To extract this information one uses tcpdump utilities developed with the standard of LBNL's Network Research Group [10]. This information is used to analyze network traffic to find a signature of abnormal network behavior and to detect possible intrusions.

The important difference of the proposed approach from traditionally used methods is the presentation of information contained in headers in terms of well-defined numerical functions. To do that we have developed software to read binary tcpdump files and to represent all protocol parameters as corresponding time-dependent functions. This gives us the opportunity to analyze complete information (or a chosen fraction of complete information that combines some parameters) for a given time and time window. The ability to vary the time window for the analysis is important since it makes possible extracting different scales in the time dependance of the system. Since different time scales have different sensitivities for particular modes of system behavior, the time scales could be sensitive to different methods of intrusion.

As was done in reference paper[11], we divide the protocol parameters for host-to-host communication into two separate groups with respect to the preserving or changing their values during packet propagation through the network (internet). We refer to these two groups of parameters as "dynamic" and "static". The dynamic parameters may be changed during packet propagation. For example, the "physical" address of a computer, which is the MAC parameter of the Ethernet protocol, is a dynamic parameter because it can be changed if the packet has been re-directed by a router. On the other hand, the source IP address is an example of a static parameter because its value does not change during packet propagation. To describe the information flow, we use only static parameters since they may carry intrinsic properties of the information flow and neglect the network (internet) structure. (It should be noted that the dynamic parameters may be important for study of network structure properties. Dynamic parameters will be considered separately.)

Using packets as a fundamental object for information exchange on a network and being able to describe packets in terms of functions of time for static parameters to analyze network traffic, we can apply methods developed in physics and applied mathematics to study dynamic complex systems. We present some results obtained in references [11, 12] to demonstrate the power of these methods and to recall important results for network monitoring applications.

It was shown [11] that to describe information flow on a network one can use a small number (10 - 12) of parameters. In other words, the dimension of the information flow space is less than or equal to 12 and the properties of information flow are practically independent of network structure, size and topology. To estimate the dimension of the information flow on the network one can apply the algorithm for analysis of observed chaotic data in physical systems, the algorithm suggested in paper [13] (see also ref. [14]and references therein). The main idea relates to the fact that any dynamic system with dimensionality of N can be described by a set of N differential equations of the second order in configuration space or by a set of $2N$ differential equations of first order in phase space.

Assuming that the information flow can be described in terms of ordinary differential equations (or by discrete-time evolution rules), for some unknown functions in a (parametric) phase space, one can analyze a time dependance of a given scalar parameter $s(t)$ that is related to the system dynamics. Then one can build d-dimensional vectors from the variable s as

$$y^d(n) = [s(n), s(n+T), s(n+2T), \ldots, s(n+T(d-1))] \tag{1.1}$$

at equal-distant time intervals T: $s(t) \rightarrow s(T \cdot n) \equiv s(n)$, where n is an integer number to numerate s values at different times. Now, one can calculate a number of nearest neighbors in the vicinity of each point in the vector space and plot the dependance of the number of false nearest neighbors (FNN) as a function of time. The FNN for the d-dimensional space are neighbors that move far away when we increase dimension from d to $d+1$ (see, for details ref.[11]).

The typical behavior of a scalar parameter and corresponding FNN plot are shown in Figs. (1.1) and (1.2). From the last plot one can see that the number of FNN rapidly decreases up to about 10 or 12 dimensions. After that it shows a

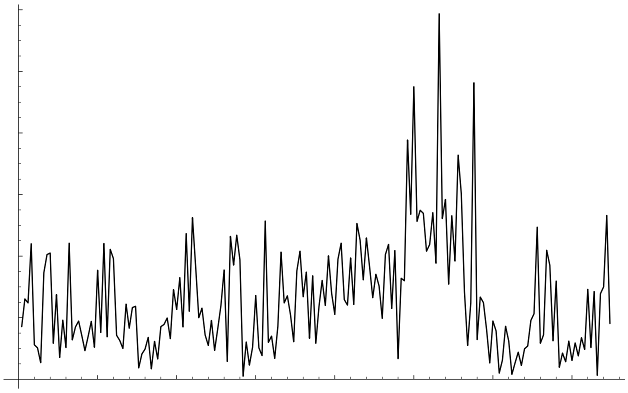


Figure 1.1: Protocol type ID in the IP protocol as a function of time (in $\tau = 5sec$ units).

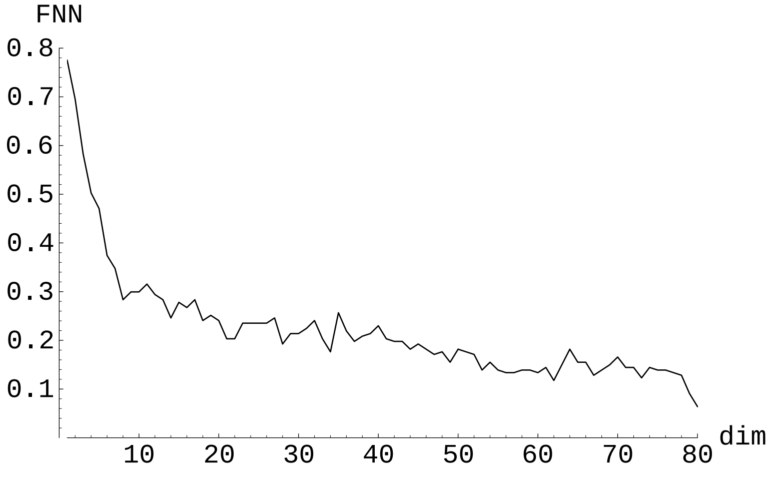


Figure 1.2: Relative number of false nearest neighbors as a function of dimension of unfolded space.

slow dependency on the dimension, if at all. Fig. (1.2) shows that by increasing the dimension d step-by-step, the number of FNN, which occur due to projection of far away parts of the trajectory in higher dimensional space is decreases to a level restricted by system noise that has infinite dimension. Therefore, for a complete description of the information flow one needs not more than 12 independent parameters. The dynamics of information flow can be described as a trajectory in a phase space with the dimension of about 10 - 12. Since this dimension does not depend on the network topology, its size, and the operating systems involved in the network, this is a universal characteristic and may be applied for any network.

However, we cannot identify exactly these independent parameters. Due to the complexity of the system it is natural that these unknown parameters which are real dynamic degrees of freedom of the system would have a complicated relationship with the parameters contained in the network protocols. Fortunately, the suggested technique provides very powerful methods to extract general information about the behavior of dynamic complex systems. For example, the obtained time dependence of only one parameter, the protocol ID shown on Fig.(1.1), is enough to reconstruct the trajectory of the information flow in its phase space. The reconstructed projection of the trajectory on 3-dimensional space is shown on Fig. (1.3). Therefore, one can see that the complete description of the network information traffic in terms of a small number of parameters is possible. The important point is that this trajectory (usually called as an "attractor") is well-localized. Therefore, it can be used for detailed analysis and pattern recognition techniques. It should be noted that the attractor presented here is obtained from one parameter measurement only, for that being illustrative purposes. For real analysis we use multi-dimensional high accuracy reconstruction.

1.3 Real Time Network Monitoring and Detection of Known Intrusions

The proposed approach for network traffic description provides the possibility of real-time network monitoring and detection of all known network attacks. This is because one collects from tcpdump binary output the complete information about network traffic at any given point in the network. All header parameters are converted into time dependant numerical functions. Therefore, each packet for host-to-host exchange corresponds to a point in the multidimensional parametric phase space. The set of these points (the trajectory) completely describes information transfer on the network. It is clear that this representation provides not only the total description of the network traffic at the given point but also a powerful tool for analysis in real time. Let us consider some possible scenarios for the analysis.

Suppose we are looking for known network intrusions. The signature of an intrusion is a special set of relationships among the header parameters. For ex-

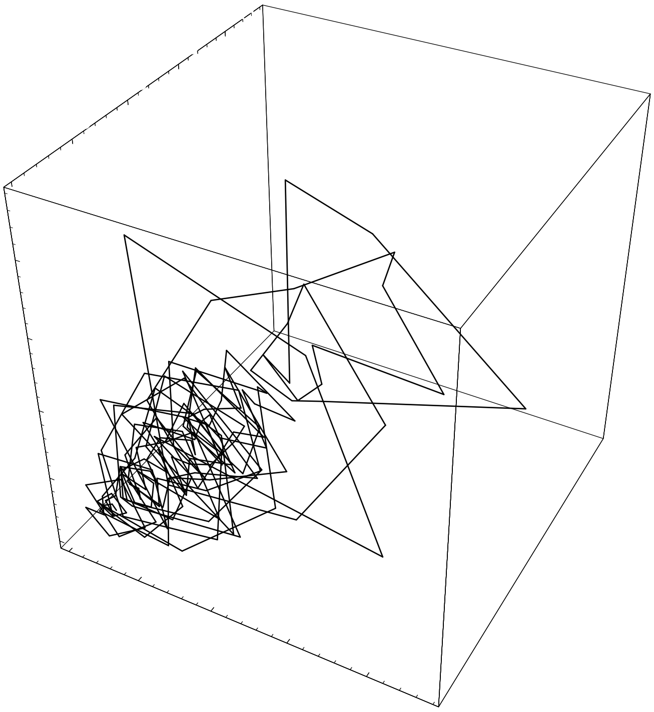


Figure 1.3: The projection of the trajectory of the information flow 3-dimensional phase space.

ample [9], the signature for the attempt to identify live hosts by those responding to the ACK scan includes a source address, an ACK and SYN flags from TCP protocol, a target address of the internal network, sequence numbers, and source and destination port numbers. The lone ACK flag set with identical source and destination ports is the signature for the ACK scan. This is because the lone ACK flag set should be found only as the final transmission of the three-way handshake, an acknowledgement of receiving data, or data that is transmitted where the entire sending buffer has not been emptied. From this example one can see that the intrusion signature could be easily formulated in terms of logic rules and corresponding equations. Then, collecting the header parameters (this is the initial phase of network monitoring) and testing sets of them against the signatures (functions in terms of the subset of the parameters) one can filter out all known intrusions. Since we can collect any set of the parameters and easily add any signature function, it provides the way for a continuous upgrading of the intrusion detection system (IDS) built on these principles. In other words, such an IDS is universal and can be used to detect all possible network intrusions by adding new filter functions or macros in the existing testing program. It is very flexible and easily upgradable. The flexibility is important and can be achieved even in existing "traditional" IDS's. What is out of scope of traditional approaches is the mathematically optimized minimization of possible false alarms and controlled sensitivity to intrusion signals. These properties are an intrinsic feature of our approach.

The important feature of the approach is the presentation of the parameters in terms of time dependant functions. This gives the opportunity to decrease as best as possible for the particular network the false alarm probability of the IDS. This can be done using sophisticated methods already developed for noise reduction in time series. Moreover, representation of the protocol parameters as numerical functions provides the opportunity for detailed mathematical analysis and for the optimization of the signal-to-noise ratio using not only time series techniques but also numerical methods for analysis of multi-dimensional functions. The combination of these methods provides the best possible way, in terms of accuracy of the algorithms and reliability of the obtained information, to detect of known intrusions in real time.

Also, the description of the information flow in terms of numerical functions gives the opportunity to monitor network traffic for different time windows without missing information and without overflowing storage facilities. One can suggest ways to do it. One example is the use of a parallel computer environment (such as low cost powerful Linux clusters) for the simultaneous analysis of the decoded binary tcpdump output. In this case the numerical functions of the header parameters being sent to different nodes of the cluster will be analyzed by each node using similar algorithms but different scales for time averaging of signals (or functions). Thus, each node has a separate time window and, therefore, is sensitive to network behavior in the particular range of time. For example, choosing time averaging scales for the nodes from microseconds to weeks, one can trace and analyze network traffic independently and simultaneously in all these

Table 1.1: The parameters involved in intrusion signatures as shown on Fig.(1.4).

Number	Protocol	Parameter	Frequency
1	IP	Destination IP Address	3
2	IP	Source IP Address	1
3	IP	Length	1
4	IP	More Fragment Flag	2
5	IP	Don't Fragment Flag	2
6	IP	Options	1
7	TCP	Source Port	1
8	TCP	Destination Port	1
9	TCP	Urgent Flag	1
10	TCP	RST Flag	1
11	TCP	ACK Flag	2
12	TCP	SYN Flag	2
13	TCP	FIN Flag	1
14	UDP	Destination Port	2
15	UDP	Source Port	1
16	ICMP	Type	2
17	ICMP	Code	2

time windows. It is worthwhile to remember that the optimal signal-to-noise ratio is achieved for each time window independently thereby providing the best possible level of information traffic analysis for the whole network. There are three obvious advantages for this approach. The first is the possibility to detect intrusions developed on different time scales simultaneously and in real time. The second is the automatic decreasing of noise from short time fluctuations for long time windows due to time averaging. This provides detailed information analysis in each time window without loss of information. At the same time, it discards all noise related information, drastically reducing the amount of information at the storage facilities. The third advantage is the possibility to use (in real time) the output from short time scale analyzed data as additional information for long time scale analysis.

To give an idea of how many parameters are used to describe signatures of currently known intrusions we use the result of the comprehensive (but probably not complete) analysis[12] of known attacks, i.e., smurf, fraggle, pingpong, ping of death, IP Fragment overlap, BrKill , land attack , SYN flood attack, TCP session hijacking, out of band bug, IP unaligned timestamp, bonk, OOB data barf, and vulnerability scans (FIN and SYN & FIN scanning). The frequencies of the parameters involved in signatures for these intrusions are shown on Fig.(1.4). The numeration of the parameters is explained in Table 1. One can see that the number of parameters used for signatures of intrusions is rather small . This fact further simplifies the procedure of the analysis.

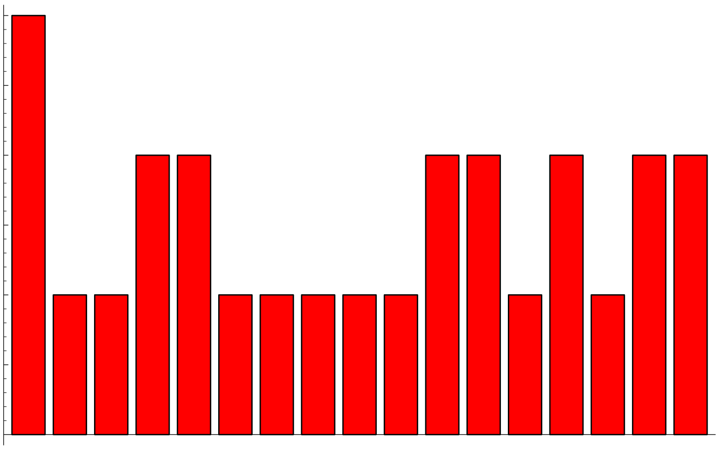


Figure 1.4: Frequencies of the parameters used in signatures of intrusions. For numbering rules see Table 1.

1.4 Detection of Unknown Intrusions

The aforementioned approach could be considered a powerful and promising method for network monitoring and detection of known network intrusions. However, the more important feature of the approach is the ability to detect previously unknown attacks on a network in a wide range of time scales. This ability is based on the method of describing information exchange on a network in terms of numerical functions of header parameters (or a trajectory in multidimensional phase space) as well as using methods of theoretical physics for the analysis of dynamics of complex systems. These methods lead to a very useful result for the small dimensionality of the information flow space. Since the number of parameters used in packet header is large (on the order of hundreds), the practical search for unknown (even very abnormal) signals would be a difficult problem, if not impossible. Therefore, the small dimension of the parametric space of the information flow is a crucial point for the practical approach for the detection of unknown intrusions.

To build a real time intrusion detection system that is capable of detecting unknown attacks, we exploit the fact that we need to analyze only a small number of parameters. Furthermore, as is known from complex systems theory, the choice of the parameters is not important unless they are sensitive to system behavior. The last statement needs to be explained in more detail. Generally, hundreds different parameters could be encapsulated in the packet headers. The question is which parameters we need to choose for the right description of the information flow. Following the discussion in the previous section, one might surmise that we need to make our choice from the known quoted 17 parameters. It may be a good guess. However, the number 17 is bigger than the dimension of the phase space which we have in mind, and it could be that hackers will invent new attacks with new signature parameters that are not included in the set presented in the previous section. The right answer to these remarks follows from complex systems theory. For a complete system description one needs only the number of parameters equal to the phase space dimension (more precisely, the smallest integer number that is larger than fractal dimension of the phase space). It could be a set of any parameters that are sensitive to the system dynamics (and the 17 discussed parameters could be good candidates). We do not know, and do not suppose to know, the real set of parameters until the theory of network information flow is developed or a reliable model for information flow description is suggested. Nevertheless, a method developed to study non-linear complex systems provides tools to extract the essential information about the system from the analysis of even a small partial set of the "sensitive" parameters. As an example, one can refer to the Fig.(1.3) which shows the 3-dimensional projection of the reconstructed trajectory from the time dependent behavior of only one parameter (the protocol ID shown on Fig.(1.1)). It means that the complete description of the network information flow could be obtained even from a small set of "sensitive" parameters.

One of the ways to implement this approach is to use the multi-window method discussed in the previous section with the proper data analysis for each time scale. This method of analysis is not within the scope of the current paper and will be reported elsewhere. We will review only the general idea and the problems related to this analysis. To detect unknown attacks (unusual network behavior) we use a deviation of signals from the normal regular network behavior. For these purposes one can use a pattern recognition technique to establish patterns for normal behavior and to measure a possible deviation from this normal behavior. However, the pattern recognition problem is quite difficult for this multidimensional analysis. According to our knowledge, it is technically impossible to achieve reliable efficiency in a pattern recognition for space with a rather large dimension, such as 10. On the other hand, the more parameters we analyze the better accuracy and reliability we can obtain. Therefore, we have to choose the optimal (compromise) solution that uses pattern recognition techniques in information flow subspaces with low dimensions. By applying appropriate constraints on some header parameters one can choose these subspaces as cross sections of the total phase space defined. In this case, we will have a reasonable ratio of signal-to-noise and will simplify the pattern recognition technique and improve its reliability. For a pattern recognition we suggest using a 2-3 dimension wavelet analysis chosen on the basis of detailed study of the information traffic on the set of networks. The wavelet approach is promising because it reduces drastically and simultaneously the computational time and memory requirements. This is important for multidimensional analysis because it can be used for an additional, effective noise reduction technique.

1.5 Conclusions

We suggest a new approach for multidimensional real time network monitoring that is based on the application of complex systems theory for information flow analysis of networks. Describing network traffic in terms of numerical time dependant functions and applying methods of theoretical physics for the study of complex systems provides a robust method for network monitoring to detect known intrusions and is promising for development of real systems to detect unknown intrusions.

To effectively apply innovative technology approaches against practical attacks it is necessary to detect and identify the attack in a reconnaissance stage. Based on new methods of data analysis and pattern recognition, we are studying a technology to build an automatic intrusion detection system. The system will be able to help maintain a high level of confidence in the protection of networks.

We thank the staff of the Advanced Solutions Group for its technical support. This work was supported by the DARPA Information Assurance and Survivability Program and is administered by the USAF Air Force Research Laboratory via grant F30602-99-2-0513, as modified.

Bibliography

[1] Réka, A., J. Hawoong and B. Albert-László, *Nature* **406** (2000), 378–381.

[2] Strogatz, S. H., *Nature* **410** (2000) 268–276.

[3] Deri, L. and S. Suin, *Computer Networks* **34** (2000), 873–880.

[4] Porras, P. A. and A. Valdes, "Live Traffic Analysis of TCP/IP Gateways", *Internet Society Symposium on Network and Distributed System Security*, San Diego, California (March 11-13, 1998).

[5] Cabrera, J. B. D., B. Ravichandram and R. K. Mehra ," Statistical Traffic Modeling for Network Intrusion Detection", *Proceedings of the International Simposium on Modeling, Ananlysis and Simulation of Computer and Telecommunication Systems*, IEEE (2000).

[6] Huisinga, T. *et al.*, *arXiv:cond-mat/0102516* (2000).

[7] Duffield, N. G. and M. Grossglauser, *IEEE/ACM Transactions on Networking* **9 No 3** (2001) 280–292.

[8] Ayedemir, M. *et al.*, *Computer Networks* **36** (2001) 169–179.

[9] Northcutt, S., J. Novak and D. McLachlan, *Network Intrusion Detection, An Analyst's Handbook*, New Riders Publishing, Indiapolis, IN (2001).

[10] LBNL's Network Research Group , *http://ee.lbl.gov/.*

[11] Gudkov, V. and J. E. Johnson, *arXiv: nlin.CD/0110008* (2001).

[12] Gudkov, V. and J. E. Johnson, *arXiv: cs.CR/0110019* (2001).

[13] Abarbanel, H. D. I., R. Brown, J. J. Sidorowich and L. Sh. Tsimring, *Rev. Mod. Phys.* **65** (1993) 1331–1392.

[14] Ecmann, J.-P. and D. Ruelle, *Rev. Mod. Phys.* **57** (1985) 617–656.

Appendix E

Patent Application
December 3, 2002
Joseph E. Johnson, PhD, Inventor &
Vladimir Gudkov, PhD Inventor

A New Methodology for the Analysis and Classification of Systems Characterized by Networks, Graphs, Clusters andTopologies.

Introduction

A network generally refers to an arrangement of objects, with connections among some pairs of the objects, but not necessarily every possible pair. Various types of networks represent the flows of people, goods, services, energy, money, power, water, and information. Transportation networks consist of highways, waterways, aircraft flight paths, or even walking trails that can be represented by lines that connect points called ‘nodes’ where the paths (roads, waterways etc) join at juncture points such as towns. Electrical networks consist of wires or conductor paths for electricity that connect electrical devices (at nodes) as in an electrical power grid distribution system, an integrated circuit or chip, or simply a device with parts wired together for electrical operation. The national and global connections among computers known as the Internet and related computer networks are examples that consist of computers, (and information processing devices such as hubs, routers, bridges, and switches) constituting the nodes connected by optical fibers and electrical connections. Communications networks include both telephone and telegraph networks. Economic networks represent the flow of goods and services in return for currency of equal value. Energy flow networks and food chain networks can represent the flow of solar and related energy among environmental components and thus could be called an ecosystem network. Various utility networks which describe both water supply and liquid (or solid) waste removal are prime examples of networks. Human (as well as animal and plant systems) contain internal biological networks for the flow of blood, fluids, nerve signals, and nutrients.

Because of the ubiquitous nature of networks, it is of the greatest importance to be able to describe the behavior of networks, and to classify their properties and to model their dynamic behavior. In the current state of the art, networks are not well understood nor can their topologies (connectivities) be classified. More specifically, we mean that there is no set of symbols or numbers that exactly correspond, in a one to one manner, to the different connectivities or topologies of networks. The methods in this patent application present a novel advance on these problems, greatly improving the current state of the art, including the implementation of these new solutions on computational devices. These methods will be shown to lead to descriptions of much more complex systems than those as represented above, and thus encompass a vast array of novel solutions to network related problems of a very fundamental importance. It is important to realize that most networks in the real world are not highly dense but consist of local clusters of high connectivity where the clusters are interconnected with only a few connections.

There are two primary purposes and areas of importance of the invention presented here. First, our invention, when embodied in computational devices, can describe complex dynamical behavior of all of the physical systems that can be described by networks including the networks listed above. The dynamical solutions can indicate weak points in the network and exactly how to correct them such as nodes and paths where flow is inhibited or reduced. Furthermore our methods predict the normal dynamical behaviors of various topologies and thus can be used to find abnormal dynamics in internet designs or detect potential network intrusions. Secondly, this invention can identify types of networks by characterizing them by a set of numbers, which classification, obtained in a computational environment or system, analyzes the intrinsic connectedness of a network and can sense if two networks are the same or different. This foundation of classification will serve as the methodology, when implemented in computers, of classifying whole topologies, local clusters, characteristic network behaviors, and component topologies.

Current Art and Mathematical Foundations

In its abstract form, we can represent a network as a set of points called ‘nodes’ which, for identification, are numbered with the integers (1,2,…n) for a system with n nodes. The connectivities among pairs of nodes are represented by lines which join nodes and can be represented by what is called in the literature as the ‘connectivity’ or ‘adjacency’ matrix: L_{ij} which is often defined to be equal to ‘1’ if nodes i and j are directly connected, and ‘0’ otherwise, where i and j range from 1 to n. The diagonal elements, L_{ii}, are traditionally set to ‘1’ if a node is considered as ‘connected to itself’ or set to ‘0’ if it is not, thus defining the complete matrix with either ‘0’s or ‘1’s on the diagonal. For certain problems the diagonal is set equal to the negative of the sum of the non-diagonal elements in that column. This results in a matrix that has the sum of all elements in each column equal to zero. As the connectivity matrix is symmetric, the sum of elements in each row will also be zero. This method of setting the diagonal is called the Lagrangian form of the connectivity matrix. In any of the three methods described above for the determination of the diagonal, it is obvious that each describes the connectivity in the same way, as connectivity is described by the off-diagonal elements. Thus to a certain extent, the method of determining the diagonal is somewhat arbitrary without other considerations.

Many different connectivity matrices describe the same ‘topology’ or connectivity among the nodes. The root of this problem is that the numbering of the nodes and the consequential assignment of the matrix elements as ‘0’ or ‘1’ is as arbitrary as the numbering. The central problem is then to devise a mathematical technique to distinguish different networks or graphs and even more generally to classify all possible graphs of a given order (number of nodes) in a unique way and thus to eliminate the arbitrariness of the node number assignment but to not discard the essential ‘connectivity’ and thus to uniquely classify the topology itself. This problem is known to be of extreme complexity and difficulty.

The current art is to manually draw all possible topologies of a given order and to visually compare a given topology to another to see if the topologies are the same. By ‘the same topology’ or the ‘same graph’ or the ‘same network’ we mean the following:

Two networks (equivalently called graphs) are topologically the same if and only if one of them can have a unique pairing of nodes of network ‘A’ to the nodes of network ‘B’, so that the connectivity matrix is the same. For example a graph where one node is connected directly to each of four other nodes (and no other connections are made) is the same whether that central node is numbered as 1, 2, 3, 4, or 5.

A number of researchers have independently suggested that the eigenvalues of the connectivity matrix (by any of the three methods of assigning the diagonal discussed above) will have values, which are in one to one (isomorphic) correspondence for, and only for, topologically identical networks. This is known to fail for each of the three methods of assigning the diagonals listed above. It is true that the resulting eigenvalues “almost” distinguish the topologies except for a small percentage of networks which are called ‘isospectral’ meaning that the same set of eigenvalues represents two different topologies. One might also say that the associated eigenvalues spectrum is ‘degenerate’ as several states correspond to it. But in the final analysis, although the connectivity matrix eigenvalue method distinguishes many of the topologies, it fails to distinguish a small percentage for n=5 nodes and higher n graphs. Many investigators have continued by studying the eigenvector components, which is equivalent to studying the angles between the nodal basis (1,2,..) and the eigenvectors associated with each eigenvalue. But this has not been very productive and cannot be automated. One is left in the current art with manual methods of visually comparing graphs and networks and exhaustingly drawing them by hand in tables or with computer programs that increase in time as the number of possible combinations thus proving impossible for any known computer in human lifetimes.

Prior Relevant Work of the Inventor

A Markov matrix is a n x n (square) matrix where n is any positive integer, with non-negative elements and with the sum of all elements in each column equaling the value ‘1’. Markov matrices have the defining property that when they multiply a vector of non-negative components, the resulting linear transformation gives a new vector that also has non-negative components and where the sum of the new components is the same as the original sum of the components. Thus Markov transformations preserve the sum of elements of a vector much like a rotation preserves the sum of the squares of components of a vector. Markov matrices do not have inverses and consequently do not form a mathematical group of transformations (because of the non-negativity condition).

However, the inventor previously discovered that by relaxing the condition of non-negativity on both the vectors and on the matrix, and retaining only the requirement that the sum of components of a vector is conserved by the transformation, that one obtains a ‘Markov-Type’ continuous (Lie) group of transformations, M, and an associated generating transformations (Lie algebra), L, where M=exp(tL). Then, by exploring the associated Lie algebra of elements ($L = a_1L_1 + a_2L_2 + a_3L_3 + \ldots + a_nL_n$ where L_k (or as defined below with two indices as L_{ij}) is the basis for the Lie algebra). The first inventor was able to show that for a certain definition of the basis L_i that all Markov transformations that are continuously connected to the identity are obtained using M=exp(tL) where the L is defined using non-negative a_i . It was also shown that the entire general linear group in n dimensions GL(n,R), can be decomposed into these

Markov type groups and diagonal transformations (an Abelian scaling group) that simply multiplies each component by an exponential growth or decay factor. The power of this result is that now one can use all the mathematical power of continuous group theory to address and classify the dynamical process of Markov transformations. It is critical here to note that the L_{ij} are defined very precisely as matrices with a '1' at a position (i,j) and '0' elsewhere except with the L_{jj} (diagonal) term which is set to '-1'. This has the result that any linear combination of the L_{ij} have the property that the sum of elements of any column are equal to zero. This is the defining characteristic for the L matrices that give all of the Markov matrices via the equation M=exp(tL).

Since Markov transformations preserve the sum of components of a vector, these transformations have been very useful over the last century for studying socioeconomic processes where for example the sum of the number of people or money is kept constant while discrete Markov transformations perform shifts from one place to another. The previous work of this inventor now allows for the study of such processes as a continuous time evolution rather than the discrete action of a Markov matrix that moves one forward in time by finite leaps. It also brings all of the power of Lie groups and algebra to provide a greater understanding of such processes as shown in prior literature.

Description of the Current Invention and Discovery

While the inventor was working on methods of describing network connectivity (in order to better characterize the behavior of different computer networks such as the Internet), he characterized the connectivity matrix of the network as being composed of '1's and '0's as had been done by other researchers as described above. He then realized that these matrices were combinations of exactly those basis elements of the Markov Lie group IF one set the diagonal elements equal to the negative of sum of the non-diagonal elements in the column. Then this resulting 'connectivity' matrix is precisely a member of the Markov Lie algebra of generators generating a Markov transformation M=exp(tL) parameterized by t and leaving the sum of the vector components conserved upon which it acts invariant. He furthermore realized that to set the diagonal elements to '0' or to '1' (or in fact to any value rather than the negative of the sum of the non-diagonal elements of the column) was equivalent to simultaneously invoking an exponential growth or decay of the conserved quantity at that node making that node a source or sink for the quantity that would have otherwise been conserved. The value of this approach by the inventor in utilizing a different interpretation for these matrices, as transformations, will lead to the deeper insights forming the basis of this patent application.

Returning now to the Lie algebra method of assigning the diagonal elements, in the study of information flow, the inventor proposes that the vector components could represent the 'amount of information' (or other entity) at each node and thus this transformation, M, provides a dynamical model for the conserved flow of information in a continuous fashion among all the nodes of a graph as in an Internet structure. This in tern led the inventor to invent this dynamical model as describing the flow of any conserved entity (information, water, energy, electricity, money, people, etc) in a network as described by the fundamental connectivity matrix. The power of this realization is that it gives a fundamental meaning and connection to the connectivity matrix both in terms of a dynamical model of network flows but also a fundamental meaning and connection to the entire theory of Lie groups and Lie algebras and specifically the decomposition of the general linear group as previously done by the inventor.

With this discovery, the inventor then studied the diagonalization of the connectivity matrix, L, and discovered that the eigenvalues were the dynamical rates at which the various linear combinations (the eigenvectors) of information were approaching equilibrium. The eigenvectors are then those linear combinations of the nodes which have a characteristic dynamical behavior much like the vibrational frequencies of normal nodes in other dynamical systems. This dynamical model proposed by the inventor now leads to a deep interpretation for networks and graphs and their expression through the connectivity matrix, via the associated dynamical models and Lie group theory. Stated in another way, the prior art represented a graph or network topology as a static connectivity matrix with little connection to other fields. With the inventor's discovery of the interpretations, analogies, and connections described above among these multiple mathematical fields, a whole new richness now unfolds that will provide vast computational depth to the understanding of both static graphs and networks and their full connection to associated dynamical flows.

Specifically, the inventor's model and analogies, lead to all of the subsequent results which, when derived from computational devices, are part of this invention: First one sees that the connectivity matrix could have diagonal elements determined by the requirement that the sum over the rows for each column is equal to zero. This makes the matrix an element of the Markov Lie algebra and it consequently generates a continuously evolving dynamical Markov transformation that conserves a transfer of information (or other conserved entity) among all the attached nodes of that network or graph as described by the connectivity matrix. Immediately from the Lie group theory the inventor showed that this connectivity matrix with '1's and '0's on the off diagonal positions describes exactly a non-directed graph or network that has equal bandwidth among all connecting pairs and thus giving exactly equal flow rates. Any other non-negative values could also be used in off-diagonal positions thereby representing more complex networks with asymmetric flows and with inter-nodal flows at different rates. The mathematical interpretation of the eigenvalues is that they are precisely the rates of decrease of these linear combinations of nodal information as represented by their associated eigenvectors. This result explicitly lets one now determine that linear combination of nodes (eigenvector) that has a unique dynamic behavior with a decay rate given by the associated eigenvalue. These results now allow a user of this system, embodied into a computer model, to model complex flows of any conserved entity with equal bandwidth connectivity as described by a given connectivity matrix, and to model the connectivity matrix by these same flows. If the diagonals are set as above so that the sum over rows of the L matrix are zero for each column, then one gets conservation of the sum of components. If instead one sets the values to be '1' or '0' or another value, this will represent as associated growth or decay rates of information at the nodes. If these systems represent symmetric flows (ie $L_{ij} = L_{ji}$) then the eigenvalues will be real. Otherwise the eigenvalues will be complex numbers and represent a more complex dynamical evolution. A specific new example of this is the description of a directed graph which allows transfer in one direction but not in the reverse direction. Such a connectivity matrix will be represented by $L_{ij} = 1$ but with $L_{ji} = 0$. Thus this invention relates also the full classification and description of directed graphs. The previous methodology describes the entirety of all possible graphs and networks that provide any bandwidth in either direction (including zero thus disallowing transfer).

The inventors first set of claims is for the representation, in computational systems, of the associated network matrices as described above, along with the dynamical representations in terms of eigenvectors and eigenvalues, for the determination of flows in networks defined by the Lie algebra. This set of claims includes but is not limited to the methods for (a) the dynamical modeling of network systems by transformations derived from the static connectivity matrix, and conversely (b) the use of the dynamical parameters, eigenvalues, eigenvectors, and methods of setting the diagonals for understanding and classifying the static properties of the graph or network, and its underlying topology. (c) The next claim presented here relates to the improvement in the description of the topology of a graph utilizing computer systems. As described above, our designation of the L matrix representation of an undirected, equal bandwidth graph as a connectivity matrix is not new in itself as it previously was utilized in another context and called the Lagrangian matrix when the diagonal terms are determined as above. It is known that the eigenvalues of L alone are not sufficient to determine the topology of the graph or network uniquely, or more significantly, than one can achieve with the eigenvalues of L when the diagonal is replaced by either ‘1’s or ‘0’s. None of the invention components or insights presented up this point are of direct help in improving the classification of the topologies of traditional networks

Specifically, the connectivity matrix shows the adjacent connectivities between neighboring nodes. In what follows, the inventor designs a basis for the analysis of higher order connectivity as a novel method for topological identification. Higher order connectivity is achieved by taking the fundamental matrix to various powers. But it is quickly realized that in the expansion M(t) = exp(tL) one has all powers of L (as it is the expansion of e^(tL) (defined by 1 + tL + ((tL)^2)/2! + …) as is necessary because M provides the continuous flow of information among all nodes. Thus there is no new information in the powers of the basic L matrix (with any form of the diagonal) as all powers are diagonalized when L itself is diagonalized. Thus any isospectral degeneracy would remain in the eigenvalue spectra of higher powers of L.

For the following novel work, we proceed with a non-linear approach. We set the diagonal matrix elements equal to zero thus defining a connectivity matrix K^1. By multiplying the matrix by itself, we obtain a matrix, $K^2 = K*K$. The resulting elements of K^2 give the number of ways that the nodes can be connected by two transitions. We remove the diagonal and place the values, in order, in the first row of a new matrix ‘S’ and then place zeros on the diagonal of K^2. These diagonal values give the number of different ways that a transition can occur from a node i back to that node in two steps and thus represents the self-connectivity in second order. We next form $K^3 = K*K^2$ and as before we remove the diagonal and place it in the second row of the matrix S and put zeros in the diagonal positions of K^3. These values give the number of different transitions that one begin at a node and return to that node only after exactly three transitions. We continue this process for 2n-2 steps which provides for exactly n-1 steps to the n^{th} node and the n-1 steps back to the original node. This process thus explores all self connecting transitions and counts those that do not revisit the node before the requisite number of steps are executed (thus is self avoiding). One notes that the 2(n-1) x n matrix S which we call the self-connectivity matrix, consists of a column vector of length 2(n-1) for each node that consists of number of unique ways that that node is self-connected. For each node, this column vector is independent of the numbering of the nodes. One notes that the process of removing the diagonal and replacing it with zeros is a non-linear process that makes the rows of S and the remaining 2(n-1) matrices K^m linearly independent. Thus it follows that the eigenvalues sets for each of the 2(n-1) of the K^m matrices are also linearly independent. .

The K^m_{ij} matrices give the number of ways that independent paths are formed that provide transitions from i to j without revisiting i in the process. These matrices 'feel out' the network in transitions of m nodes at a time. We replace the zero diagonal values with the negative of the sum of the values in that row thus making each component matrix a generator of a continuous Markov transformation. We find the eigenvalues of each of the K^m_{ij} matrices and use the n eigenvalues to form a row of a new matrix, 'E' which will have 2(n-1) rows corresponding to the eigenvalues of K^m_{ij} . The eigenvalue sets for each m represent the rates of approach to equilibrium for the associated eigenvectors of the transition matrix of that order. The E matrix is totally independent of the node labeling. Next we take all of the n eigenvalues for each value m = 1 to 2(n-1) and construct a new matrix 'V' by placing the eigenvalues in rows in order by the value of the eigenvalue and next in order the value of m. This gives us a matrix with the columns labeled with the node numbers (as with S) and the rows labeled with the eigenvalues values at successively higher values of m.

We now have three matrices S, E, and V of which the columns of S and V depend upon the node numbering. By adjoining S to the top of V, the numbering of the nodes is maintained. Successively, each column of this adjoined matrix is ordered by sorting the columns in order of the values in the first row, then the second only reordering those columns which have up to that point are not uniquely ordered. This methodology provides unique and extensive metrics for the description of many of the aspects of the network and associated clusters as delineated below. We note that the derivation of the matrices S, E, and V is lossless in the sense that one can retrieve all information of L_{ij} from these matrices. We also note that the algorithms for the derivation of these matrices are extremely fast and increase linearly with the number of nodes. In particular, the determination of S is extremely fast even for very large networks.

Additionally, to identify, simplify and accelerate the algorithms for graph and cluster identification, we have invented still additional tools and metrics in the form of algorithms that the measure of the identity of the connectivity matrixes, their products and their internal structures (topology, order etc). These functions are refered to as the mutual information (or entropy) of the S, V, and E matrices including specifically the connectivity matrix itself. These measures give unique and invariant information about network topology represented by a set of real numbers called generalized entropies. We here describe the algorithm for the calculation of mutual entropies by considering a matrix C which could be any of the matrices discussed above.

Let a graph or network G which consists of vertices V_i connected by edges E_{ij}. It is described by a connectivity matrix C with $C_{ij} = 1$ if i is connected to j, else $C_{ij} = 0$. We also set $C_{ii} = 0$. A vertex relabeling $i \to p(i)$ leaves G invariant but changes C according to

$$C \to C' = P^T CP \qquad (1)$$

with P an orthogonal matrix with only one non-zero element in each row i and column $j = p(i)$, which represents the above permutation

$$P = \delta_{j,p(i)}. \tag{2}$$

If we normalize the connectivity matrix C so that

$$\sum_{i,j=1}^{n} C_{ij} = 1, \tag{3}$$

then $P_i = \sum_j^n C_{ij}$ could then be considered as the probability that V_i and V_j are connected. The Shannon entropy, (corresponding generalized entropies could be used instead of the Shannon one),

$$H(row) = -\sum_{j=1}^{n} P_i \log P_i \tag{4}$$

is a measure of the uncertainty of the connections for a given network. The amount of *mutual information* gained via the given connectivity of the network is

$$\begin{aligned} I(C) &= H(row) + H(column) - H(column \mid row) \\ &= \sum_{i,j}^{n} C_{ij} \log(C_{ij} / P_i P_j), \end{aligned} \tag{5}$$

where

$$H(column \mid row) = -\sum_{i,j}^{n} C_{ij} \log(C_{ij}). \tag{6}$$

$I(C)$ does not depend on the vertex relabeling and is a permutation invariant measure for the connectivity matrix. If the mutual entropy for two connectivity matrices are different, they correspond to different graphs. We found that the entropy is already sufficient to distinguish even between graphs that are normally cospectral.

The extension of the algorithm for calculation of Shannon entropy could be used for calculations of mutual Rényi entropy and Tsallis entropy. For example, based on definition of Rényi entropy, the expressions in eqs.(5) and (6) are transformed into

$$H_q(row) = -\frac{1}{1-q} \log \sum_{j=1}^{n} P_i^q$$

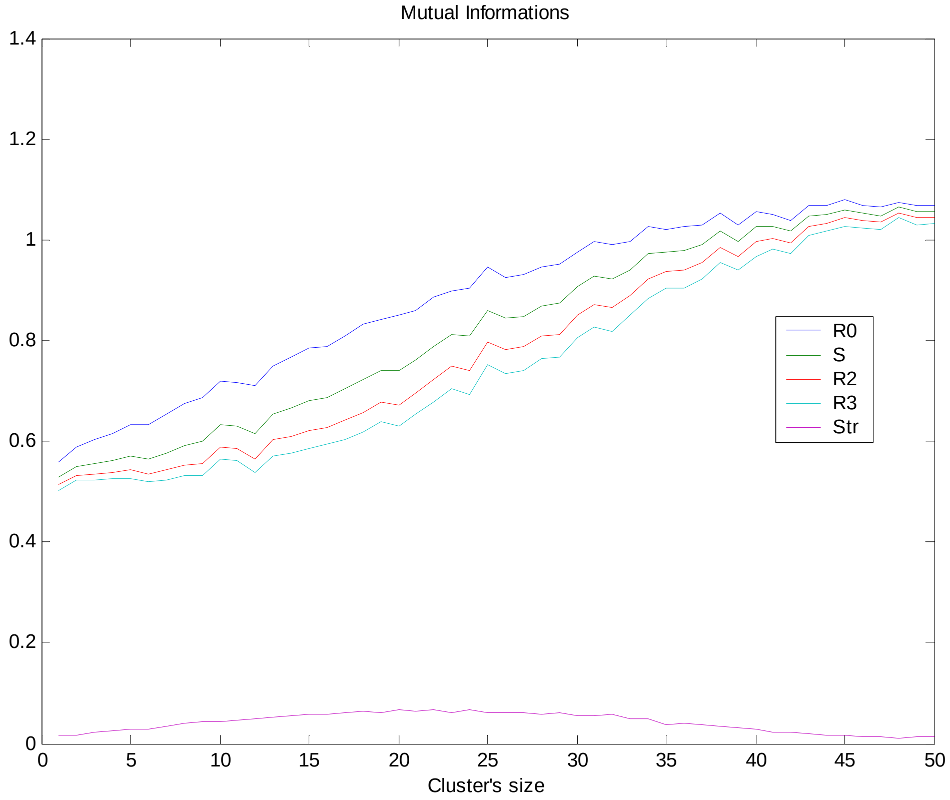
and

$$H_q(column \mid row) = -\frac{1}{1-q} \sum_{i,j}^{n} \log(C^q{}_{ij}),$$

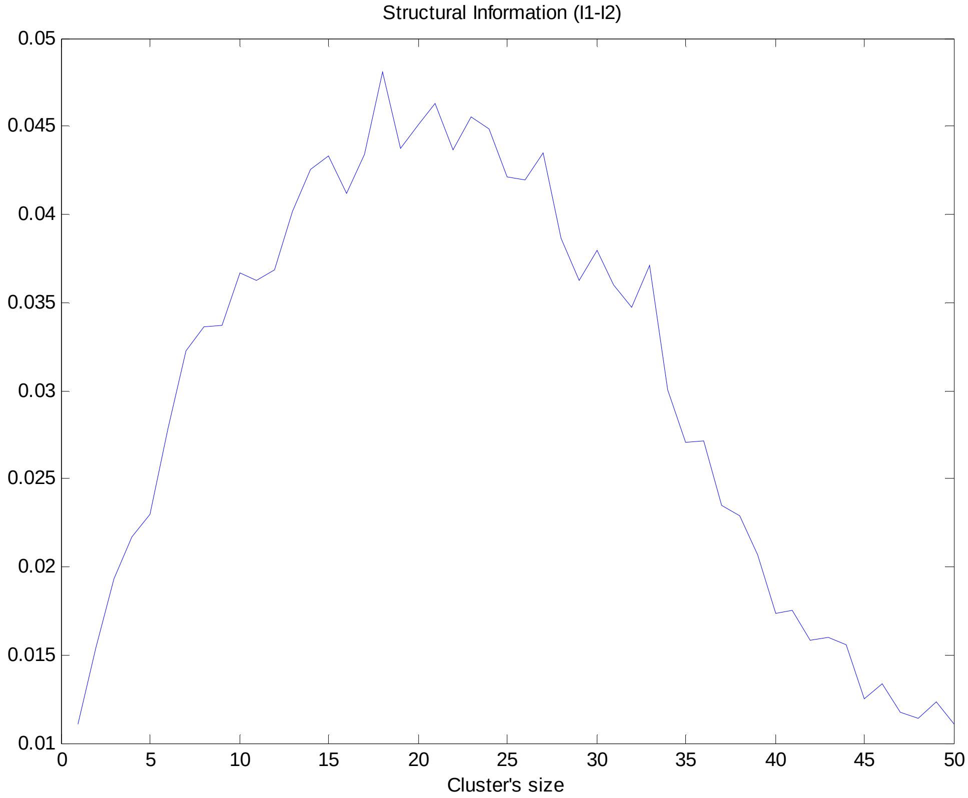
giving mutual information Rényi $I_q(C)$ for the given matrix C.

Using the Rényi information (and, for essentially non-equilibrium network dynamics cases, Tsallis information), one can not only distinguish between different network topologies on the base of the connectivity matrixes but extract information about network topology, such as number of clusters, cluster's dimensionalities etc. Moreover, by monitoring appropriate functions of mutual information, one can observe in real time a change in topology of the given network including a cluster's formation, disappearance or appearance of group’s connections, change of the connection “styles”, and other features.

For example, during the formation of a new cluster with a high level of connectivity in a 100 node network, Shannon and Rényi mutual information smoothly increase with the size of the cluster as shown on the Figure 1.



However, a difference of mutual Shannon information and Rényi information of kind 2 (q=2) displays the sharp dependence of the size of the formed cluster (Figure 2).



This gives a unique opportunity to monitor dynamical behavior of the network in real time. It should be noted that different entropies are sensitive to different patterns of network topology (such a size of clusters, number of clusters, fractional dimensionality, etc), therefore many important properties of network can be extracted using suggested methods.

In the following, our claims refer to the mathematical functions and matrices defined above, as embodied in computational devices, whether the device is mechanical, biological, electrical, optical, or any combination of these in either analogue or digital embodiments. The specific software algorithms for these functions and matrices are known to anyone skilled in this art. Specific claims and applications follow:

Claims

1. The function $I_q(C)$ can be used in most cases to uniquely distinguish network (graph) topologies, and can be adjoined to the S, E, and V matrices to provide a vast depth of information not just to distinguish a network (or subnetwork) but to provide useful knowledge on the statics and dynamics of the clusters and their hierarchies. The value of this claim is even greater due to the extremely fast ability of these functions to be calculated for large networks and graphs.
2. As larger network can be viewed as a network (joining) of smaller networks it follows that larger structures can be approximated as networks of smaller

structures. The embodiment of the functions in equations (5) – (7) for the identifying of networks and sub-networks is clamed as a novel method when embodied in any computational system.

3. To identify internal network structures and its dynamical behavior (changes of the structure as a function of some parameters) the equations for the Shannon entropy (4) – (7) can be replaced by the corresponding sets of Renyi and Tsallis entropies. (Generalizations and combinations of these entropies are used in the claimed algorithm to identify the specific feature of networks / graphs.)
4. General mutual information can be used for calculation and monitoring of survivability and assurance of network stability for distributed forces on a battle field and in distributed business operations.
5. Appropriate functions of $I_q(C)$ can be used for automatic calculation of the level of stability and proper functionality of dispersed computational and/or remote sensors network for self-regulation of the level of connectivity with a given assurance for survival.
6. Monitoring of the set of functions $I_q(C)$ can detect of unusual activities (and formation of new groups of collaborators) on a network (communication, material supply, etc).
7. Real time monitoring of proper $I_q(C)$ can be used for determination in advance possible outrages and weak chains in the power supply, resources distribution and communication networks.
8. Calculation of $I_q(C)$ for large data bases (data warehouses) with diverse sources of incoming data can be used for identification and localization of clusters of correlated data sets.
9. The self connectivity matrix, S, defined above provides the a novel spectrum of the clustering associated with any node and at all chain links. These spectral values can be compressed by weighting with a exponentially decreasing function of the node level (or similar measure) to give a single numerical measure of the cluster size associated with each node.
10. The eigenvalues contained in E give the rate of approach to equilibrium for systems described by the associated connectivity matrix from which it is derived and for the nodal weights as expressed in the associated eigenvector in V.
11. Continuous transformations as generated by the component 2(n-1) different generating matrices contained in V provide transformations that are non-revisiting for the initializing node.
12. Specifically we claim the application of these algorithms to the monitoring of both internet and telephone networks for the detection of cyberterrorism, unauthorized intrusions and use as well as for the identification of aberrant behavior indicating a system malfunction.
13. Specifically we also claim the application of these algorithms, and specifically the use of information metrics, to the identification of the survivability of distributed (networked) systems in military and battlefield deployments as well as the distributed network channels for business and transportation monitoring.

Applications

The very novel methods described in this patent application, when implemented in computational devices as described above provide very substantial inventions in two domains: (1) The classification, organization, and identification flows of information, money, goods, people, electricity, water, and other quantities listed above, of dynamical behaviors of all manner of networks thus providing the methods for identifying problems, improving flows, and detecting both intentional intrusions and natural failures. (2) The classification and full description of topological structures including the critical area of network cluster identification as associated with all forms of networks and the classification of subnetworks as identified by the numerical signatures in the S, E, V, and Information functions and matrices described above. Of the greatest importance and novelty is the extreme speed with which these algorithms work for very large clusters thus allowing both static computations of clusters and topological metrics as well as the rapid recomputation thus allowing the dynamical tracking of topological network dynamics.